Handbook of Pupil Evaluation

PRITAM SINGH

This *Handbook of Pupil Evaluation* is written in the Indian context but in accordance with the modern view of evaluation as a growth oriented ecological concept, using the philosophy of objective based evaluation which is reflected even in organisations of its content. Contemporary concepts, process and tools of evaluation are discussed to emphasize evaluation as an integral part of the teaching learning process with a focus on improvement of students learning. The book is intended for practical use of pupil teachers, inservice teachers, teacher-educators, paper-setters, examiners and evaluators. Teachers would find a wide variety of examples testing different abilities using verbal and non-verbal media. Basic statistical concepts are incorporated to facilitate conceptual understanding and efficient use of evaluation data. Detailed discussion of taxonomy of educational objectives, comprehensive exposition of tools and techniques of evaluation and exhaustive treatment of evaluation data makes this compendium complete in itself. Emphases in evaluation as reflected in National Education Policy and the emerging demands on today's evaluators are highlighted towards the end and forms the basis for taking up more complex and technical issues of measurement in the forthcoming companion volume of this publication. Topics like question banking, continuous comprehensive evaluation, diagnostic evaluation, criterion-referenced testing, illuminative evaluation, grading and scaling, practical examinations, measuring the affect, evaluation of curricula, textbooks, programmes and institutions, self evaluation, open book examinations etc. will be taken up for discussion in that volume.

ISBN 81-7023-203-1

Rs.180.00

Handbook of Pupil Evaluation

HANDBOOK OF PUPIL EVALUATION

PRITAM SINGH

M.Sc., M.Ed. (Gold Medalist), Ph.D. Associate (London University)
Professor and Head, Navodaya Vidyalaya Cell (NCERT)

Formerly
Head Department of Measurement and Evaluation and
National Talent Search Unit NCERT, New Delhi

S.C.E.R.T. West Bengal
Library
Calcutta
25/3, S. C. Road

ALLIED PUBLISHERS LIMITED

NEW DELHI BOMBAY CALCUTTA MADRAS NAGPUR
AHMEDABAD BANGALORE HYDERABAD LUCKNOW

ALLIED PUBLISHERS LIMITED

Regd. Off.: 15 J.N. Heredia Marg, Ballard Estate, Bombay 400038
Prarthna Flats (1st Floor), Navrangpura, Ahmedabad 380009
3-5-1129 Kachiguda Cross Road, Hyderabad 500027
60 Bajaj Nagar, Central Bazar Road, Nagpur 440010
16A Ashok Marg, Patiala House, Lucknow 226001
5th Main Road, Gandhinagar, Bangalore 560009
17 Chittaranjan Avenue, Calcutta 700072
13/14 Asaf Ali Road, New Delhi 110002
751 Anna Salai, Madras 600002

S.C.E.R.T., West Bengal
Date 29.1.91
Acc No. 4840

371.26
SIN

First published 1989
© Allied Publishers Limited, 1988
ISBN 81-7023-203-1

PRINTED IN INDIA
by V.K. Sharma for Veronica Press, 81-U.B. Jawahar Nagar, Delhi 110007
and published by R.N. Sachdev for Allied Publishers Limited,
13/14 Asaf Ali Road, New Delhi 110002

To
the memory of my beloved son
NAVDEEP
who did not live to see this book

ACKNOWLEDGEMENTS

It gives me not only great satisfaction but also the privilege to express my sense of gratitude to the eminent educationist Professor P.N. Kirpal formerly Deputy Chief Executive of UNESCO and Secretary Education, Government of India, for his painstaking efforts in going through this book and giving his expert opinion in the form of a very concise, thoughtful and expert opinion about this publication. My grateful thanks are due to Dr. M.B. Buch, the well-known educationist, researcher and academician, formerly Director of Centre for Advance Studies in Education, Baroda, and Head, earstwhile Department of Field Services of the NCERT, for reflecting his views about this publication especially highlighting its usefulness for teachers, teacher-educators, paper-setters, moderators and researchers.

I am thankful to Dr. (Mrs.) Kamla Menon, Lecturer, NCERT by whose courtesy one of her sample unit tests in Geography, *Man And Food Supply* was included in this book. For this I also acknowledge the NCERT for having used this material developed in the Department of Measurement and Evaluation. I would like to mention particularly the names of eminent measurement specialists like Lindquist, Thorndike, Hagen, Guilford, Ebel, TenBrink, Hudson and Popham from whose writings I culled out various ideas and quoted in this book. Lastly, I would like to offer my apologies to all those whose contributions and publications might have been taken advantage of, in one form or the other, but could not be acknowledged by name. My special acknowledgements are due to the NCERT and other educational agencies which provided me the needed opportunity to develop, grow and improve my academic acumen on the basis of which I could accomplish the task of writing this book.

PRITAM SINGH

FOREWORD

Examinations, at all levels of education, from the school to the university and professional colleges, have to be taken by students for admission, promotion and certification of achievement. For purposes of employment, and particularly for employment in the public services, one has to sit in examinations. Hence, the importance of examination is obvious.

With developments in the fields of education, training, placement, personnel management and productivity, it has become increasingly clear to all those who are concerned with instruction, management of learning, students and learners in general in various kinds of activities, that examinations are basically concerned with value judgements, expressed in the form of pass-fail, marks, grades and the like, regarding what has been learned or acquired by the individual concerned. Evaluation, instead of examination, seems to indicate better what goes on in such situations. The purpose of any examination in the school or college, be it tutorial, periodical, terminal or half-yearly, is to evaluate what has been learned, how much has been learned and how well it has been learned. An evaluative judgement is expressed in the form of marks out of 10, 15, or any other set of marks for each question, with a total of 100 marks distributed more or less equally among a fixed number of questions to be answered, say 10 or 5 within a fixed period of time.

During the last thirty years, there has been considerable development in evaluation of pupil learning. A large variety of concepts are in use, and for any teacher and practitioner of pupil evaluation, it is necessary to know all these in order to take appropriate decisions, according to the situation, the individuals concerned the purpose of evaluation and so on. In other words, as a result of research, during the last fifty years,

we know now various factors which influence evaluation and, therefore, for any worthwhile decision, we need to discriminate among the various types of evaluation and select the one which is most appropriate for the task in hand.

For any evaluation, it is now necessary to prepare a scheme first showing the specific educational objectives, the types of items for each objective and the number of items in each type. This is the beginning of objectivity of an evaluation, because there must be a consensus among several evaluators on the scheme of evaluation, including the rules of assignment of marks. The items are usually of the objective type, which in practice assumes the form of multiple-choice items with one correct answer, or best answer, as the task for the student to identify. But there are other types of items also. The important thing is to have a number of items and reduce the task of handwriting to a minimum. This is another aspect of the objectivity of a test or examination. Between examiners variations are reduced to the minimum, practically to zero. But in order to prepare an instrument based on a scheme like the one visualised here takes time and effort.

This *Handbook of Pupil Evaluation* focusses on pupil evaluation by classroom teachers in the Indian context, using the philosophy of objective-based evaluation and putting the same into practice in its content, organisation and exposition. The author Dr. Pritam Singh has the classroom teacher, the pupil teacher and the in-service teacher educator in view. Written in a simple and direct style, this handbook is based on a philosophy of objective based instruction and evaluation. In addition, evaluation is treated as a growth-oriented, total chool concept and a feedback device that aims at improving the educational process.

The author has emphasised evaluation as an integral aspect of the instructional process, with a focus on improvement of students' learning. The philosophy of objective-based evaluation is reflected in the organisation of the book. The major learning outcomes expected are highlighted in the preface, besides stating the Intended Learning Outcomes at the beginning of each chapter, to provide direction to and focu the attention of the readers on intended learning.

The first three chapters are on Historical background of measurement, Contemporary concepts and Basic statistical concepts and are useful for developing an insight into the changing emphases in testing, for better communicability among evaluators, for a better conceptual understanding and for a better use of evaluation data in improving classroom learning. There is a chapter on Taxonomies of educational objectives. A comprehensive treatment is given to various taxonomies in cognitive, affective and psychomotor domains to reflect the different frameworks for teaching and evaluation.

The Handbook is intended to be of practical use. The teachers would find a number of examples for testing different abilities, based on various content elements, using verbal and non-verbal media of testing, with the help of different forms of questions. In fact, quite a few innovative varieties of items are introduced in chapters on construction of various types of items. The technique of developing unit tests and question papers is discussed in the Indian context and changing emphasis in designing and developing question papers in various examining agencies.

It is good to see that in this handbook, the author, Dr. Pritam Singh, who has a number of years of experience in the NCERT in relation to evaluation of learning outcomes, has delineated in clear terms the various aspects of evaluation for the teacher and the teacher educator. He has plans, I understand, to produce another book on the subject concerned with some other aspects of evaluation, which have not been covered deliberately in the present volume.

I would like to congratulate Dr. Pritam Singh for undertaking this work of exposition of the concepts and methods of evaluation for the benefit of teachers in a simple and straightforward manner, highlighting the Indian context. It is hoped that the clientele of the handbook would find it rewarding in making evaluation a useful device for not only measuring students' achievement, but also for improving students' learning.

New Delhi
Date: 15.1.88

SHIB K. MITRA
Formerly Director NCERT

PREFACE

This *Handbook of Pupil Evaluation* was visualised in two volumes, one dealing with fundamentals of measurement and process of evaluation while the other one which is under print, to incorporate the more technical and complex concepts relating to emerging trends in educational evaluation. Although originally planned as one volume, compulsion of pages and the two types of clientele prompted me to bifurcate the content into two; one primarily for teachers and students of teacher training colleges and the other one for teacher educators and professional readers. Accordingly, the first volume was restricted to concept, process and techniques of evaluation besides reflecting changes in emphases and the emerging trends. The focus of this handbook is on pupil evaluation primarily concerning cognitive domain, the other two domains being taken up in the second volume.

While there is no derth of foreign books on measurement and evaluation the fact remains that there is hardly any book for Indian teacher or student teacher, which gives comprehensive view of evaluation and at the same time reflects modern approaches to depict evaluation as a total school concept, a growth oriented concept and an ecological concept in the context of educational process. A deliberate attempt is made to focus the attention of the reader on the intended outcomes of each chapter so that he may be able to look for the relevance of the material in each chapter, to the intended learning. This is indeed the basic philosophy underlying evaluation process which envisages formulation of instructional objectives and taking cognizance of teaching and testing. This makes the reading of each chapter more meaningful and provides the needed direction. For the interested reader each chapter is documented and the references made in the text are given chapter-wise at the end of the book. This I am sure would

encourage further reading besides providing authenticity of the material quoted or referred to in the text.

There are 19 chapters in this book starting with historical background of measurement and ending up with emerging demands on tomorrow's evaluator. The first chapter though, has no relevance to the educational process yet gives a backdrop of history of the field of measurement. It provides the readers an insight into various efforts made ranging from earlier crude attempts to the modern approaches with well developed theoretical constructs, thereby reflecting a trend towards precision, comprehensiveness and psychometric approach to measurement, with a focus on ecology of evaluation.

Due to a lot of literature coming up in this field a web of evaluation terminology has developed. It has caused so much confusion due to overlapping of terms that for ordinary teachers it is difficult to see relationships among these concepts and judge their scope and implications. It was, therefore, necessary to acquaint the readers with the contemporary concepts in the field of educational evaluation. Likewise assuming that most of the teachers may not be conversant with basic statistical concepts, a chapter on use of some basic concepts in statistics was considered a necessity to enable the readers to understand and appreciate the discussion made in the chapters involving statistical concepts.

Need for a thorough discussion of instructional objectives in any book on evaluation cannot be over-emphasised. Since classification of goals and their role in teaching and testing is a field by itself, their formulation, derivation, specification and statement was necessary in order to highlight objective based evaluation vis-a-vis objective based instruction. This chapter paves the way for evaluating learning outcomes and forms the basis for objective cente ed instruction and objective based evaluation.

Before discussing the process in the form of sequential steps it was necessary to dispel some misconception about the nature, scope and function of evaluation in the light of modern view of evaluation and its place in the teaching learning process. Besides, differentiating the various general characteristics emphasis is laid on treating evaluation as a growth oriented,

total school concept concerned with quality control and diagnostic functions. This is followed by a chapter on evaluative steps involving information gathering, judgement forming and decision taking.

Chapter on Tools and Techniques forms the crucial content of this handbook. It discusses that four basic techniques of evaluation, viz., testing, observation, inquiry and analysis with reference to nature, purpose, scope and tools involved in each of these techniques. Test construction being the nucleus of educational evaluation, separate chapters are devoted to the discussion on genesis of a question, constructed questions (long answer and short answer), Multiple choice items and other varieties of objective type questions. These chapters are very well illustrated with examples testing different abilities.

Forms of questions is followed by technique of construction of achievement tests. For this a separate chapter is devoted to the theory of measurement discussing in details the three basic qualities of measuring instruments, viz. validity, reliability and usability. This is followed by a separate chapter each on unit testing, and paper setting. Designing, developing blue print, question framing, consolidation, developing marking schemes and question-wise analysis are described besides moderation techniques. How validity, reliability and usability can be taken care of in such achievement tests is indicated besides providing rationale underlying designing of these tests. Analysis of questions and question papers forms the next chapter that deals with distractor analysis, facility value and discrimination value of items in a norm-referenced as well as criterion-referenced tests besides other statistical concepts related to measurement data. Simple techniques of analysis are discussed to enable the teachers to use analysis for improvement of questions and question papers.

In order to enable the readers to become aware of the changing emphases in examination reform and educational evaluation, trends are highlighted in a separate chapter. These trends relate to concept, purpose, focus, scope, approach, instrumentation, methodology of judgement making and decision making aspects of evaluation. Purpose is to make them conscious of the need for using evaluation not only for

measurement of students' achievement but also for improvement of their achievement.

The penultimate chapter which is of informative nature rather than reflective. It describes the place of evaluation and its role as reflected in the National Education Policy and the corresponding Programme of Action. It was added later to highlight the priorities in educational evaluation. This chapter lists recommendations at school as well as college/university level and provides direction for future work in this field.

The last chapter on Emerging Demands on Today's Evaluators is specifically included to identify the priority areas and emerging trends in philosophy, psychology and sociology of educational process which have corresponding demands on pupils, teachers and educational administrators for evaluating students, teachers and institutions. These learner-based, teacher-based and administrator-based demands are to be met by today's evaluators who have to plan, develop and organise evaluation process in a way that it meets requirements of the clientele. These demands are crystallised in this chapter in the form of indicators like, the need for diagnostic testing, question banking, criterion referenced testing, continuous comprehensive evaluation, scaling and grading, improvement of practical tests, measurement of affective outcomes, curriculum evaluation, institutional evaluation, needed research etc. etc. All these issues are discussed in the forthcoming volume which is under print.

This handbook can best be used by our clientele if while reading or while teaching this subject, keep in view the following main themes of this book, which are indicated at best by means of a list of the major intended learning outcomes that should result from its use:

(1) Recall the history of educational measurement as well as the nature and scope of various terms and concepts used in educational evaluation, from time to time.
(2) Understand the modern concept, process and steps involved in educational evaluation.
(3) Identify inter-relationship of teaching, learning and evaluation.
(4) Formulate, classify, specify and state instructional objectives in performance terms.

(5) Calculate various statistics and use them in interpreting evaluation data.

(6) Apply the concept of validity, reliability and usability in construction, selection, interpretation and use of tests and other tools of evaluation.

(7) Develop the skill to construct, select and use various type of questions and tests.

(8) Use non-testing techniques of evaluation in appraising student's learning and growth.

(9) Analyse and interpret tests and test results.

(10) Appreciate the potentialities and limitations of various evaluation procedures used in schools.

(11) Get insight into the role of evaluation as a quality control device and in improvement of the teaching learning process.

(12) Become aware of the modern trends, priorities and changing emphases in educational evaluation.

While writing this book students of teaching and practicing teachers are presumed not to have had any previous special training in educational measurement. That is why no attempt is made in this volume on complex and highly technical issues of measurements. It is written for those who are directly involved in teaching, paper setting, examining, testing and interpretation of test results of either external or internal examinations. As such teachers, pupil teachers, teacher educators, paper setters, evaluators and examiners are the clientele for whom practical guidelines are provided besides giving the relevant concepts and principles underlying, besides, philosophical, rational and empirical bases for understanding and accepting the suggestions offered.

There is hardly any doubt that little that is written in this handbook has not been said before. Author's best hope is that he has used his experience of 25 years in this field, as a guide for selection of material from various sources and presenting it in a manner that it becomes meaningful to the readers. Most of

the illustrative materials included in this book are developed and used by the author in various training programmes conducted by him at state, national and international levels. Nevertheless, it was quite impossible to discuss the various topics to the desired extent because of the constraint imposed on number of pages.

I take this opportunity to acknowledge my grateful thanks to all my colleagues from N.C.E.R.T. with whom I had the privilege to interact, conduct evaluation workshops and training programmes that provided me more and more insight into the development of materials and the related concepts.

For many of the new ideas and concepts discussed, I owe my sense of gratitude to measurement experts like those of B.S. Bloom, Walker Hill, Krathwhol, Arie Lewy, Klopffer, A.E. Harper (Jr.), S.K. Mitra, R.N. Mehrotra, R.H. Dave, P.N. Dave and many other experts from whose lectures, discussions and books I had my professional growth in the field of educational evaluation and ventured to write this book.

I am indebted to Dr. S.K. Mitra for having agreed to go through the manuscript and write a foreword for the same. It was indeed a great task for him and a pleasant satisfaction for me to have a foreword from a great psychometrician and an eminent educationist.

I would like to acknowledge the contribution of my wife Rajwant who actually prompted me to write this book and continually reminded me of the delay in the time targetted schedule of writing whenever I lapsed into long spell of break during writing of the manuscript.

It is my fond hope that this Handbook of Pupil Evaluation will be useful to all teachers, prospective and inservice and the teacher educators especially from the point of view of practical guidelines provided for improving evaluation practices in their institutions Besides, it is hoped that the paper setters, examiners and moderators connected with public examinations would get an insight into the rationale underlying technology of paper setting, moderation procedures and marking schemes. More significant contribution of this volume would be if my clientele realise that evaluation is an integral part of the teaching learning process and can be used not only for measuring students' achievement but also as a diagnostic and quality

control device to improve their learning.

> To know what a student knows or knows not is only measurement of his achievement but to know why he knows not and how he can know what he knows not, is improvement of his achievements.

New Delhi
13th April 1988

PRITAM SINGH
Professor
National Institute of Education
N.C.E.R.T., New Delhi

CONTENTS

Foreword

Preface ix

I. Examination Reform in Retrospect 1

II. Contemporary Concepts in Educational Evaluation 31

III. Basic Statistical Concepts in Measurement 64

IV. Instructional Objectives and Evaluation 104

V. Modern Evaluation—A Total School Concept 145

VI. Steps in Evaluative Process 164

VII. Tools and Techniques of Evaluation 187

VIII. Genesis of a Question 207

IX. Constructed Questions (Supply Type) 217

X. Objective-Type Items (Multiple Choice Type) 246

XI. Objective-Type Items (Other Varieties) 282

XII. Qualities of a Good Measuring Instrument 323

XIII. Unit Testing—A Developmental Approach 352

XIV. Technology of Setting Better Question Papers 382

XV. Moderation of a Question Paper 417

XVI. Analysis for Improving Questions and Question Papers 428

XVII. Evaluation Trends in 10+2 Pattern of Schooling 444

XVIII. Evaluation in National Policy on Education —1986 463

XIX. Emergent Demands on Today's Evaluator 474

References 489

Index 501

CHAPTER I

EXAMINATION REFORM IN RETROSPECT

1. Origin of Examinations

Examinations have played differential roles in the strategy of educational programmes. In fact they have played a far more influential role in human history than is generally recognised. Examinations, which indeed occupied the position of measurement and evaluation, originated in antiquity. Since then an examination has been considered as a judgemental balance, a sacred cow, a Socratic quiz, an infallible instrument of evaluation, a necessary evil, a yardstick for measuring the efficiency of teachers, a standard setter, a learner's goal, parents' focus and a societal criterion of employment. Their prominent position in the social millieu, key role in educational set-up, strangulating influence on the teaching-learning process and omnibus use in certification, selection, learning, teaching and employment have made examinations a prestigious institution. Their role at different times can be appreciated if we trace the origin and history of these examinations from their earliest use to the present day.

Perhaps the first evidence of oral examinations is found in a story told in the Old Testament (1) of the pronunciation test the Gileadites gave to their enemies, the Ephraimites, who had escaped to cross the Jordan. In order to recognise an Ephraimite he was asked by the men of Gilead if he was an Ephraimite. To escape arrest if he said, "Nay"; then he was asked to say, "Shibboleth". If he pronounced it as "Sibboleth", he was slewn at the passage of the Jordan. This was the one-question

test of the objective type which forty two thousand Ephraimites failed to pass and therefore, were killed. It is a different thing that the use of measurement was made for a different purpose depending upon the goal or value of measurement. If some of the Ephraimites pronounced correctly or the tester himself could not discriminate between the actual and the intended pronunciation, the sword of judgement might have fallen on the wrong man. This is a question of the relevance and dependability of the testing device that is a separate issue.

Socrates (469?-399 B.C.) used oral quizzing by putting searching questions. Job interviews are nothing but oral tests. Up to the middle of the 19th century oral examinations were used in American schools as a standard. Some countries still require oral final examinations by law (Stanley, 1960). 'Thesis oral' in American Universities is a vestigial form of this.

Written tests are comparatively of more recent origin than oral testing or quizzing but even they date back to 2200 B.C. when the Chinese are believed to have had a national system of examinations to select their public servants. Confined to isolated cells, the candidates were assigned topics on which a long paper or treatise was to be written. Recognition of individual difference, 4 B.C., led Plato to divide his ideal society into three classes, the workers, the protectors and the rulers. An impressionistic judgement of personality on the basis of physiognomy, body build or divination led to classes of persons specialising in astrology, palmistry, phrenology and graphology (2).

The first educational tests were like present performance tests in that they did measure physical performance. In primitive tribes, young men were tested for their knowledge of tribal customs, endurance, bravery, hunting skills etc. needed for protection. The ancient Spartans (3) conducted examinations as early as 500 B.C. to test physical development and stoicism which were stressed in their educational curricula. Evaluation of athletics and aesthetic development by means of games and contests in reading, writing and singing in ancient Athens (4) is another example of the use of tests. In medieval times; oral examinations were used in the University of Bologna by 1219 A.D.

and the University of Paris by the end of 13th century when candidates were required to defend their thesis orally. Perhaps for the first time educational use of written examinations was made at Cambridge, England, in 1702 (5).

2. Developments during 19th Century

One of the earliest records of educational testing in America dates back to the written examination in Boston in the year 1845 (6). Prior to that, oral examinations prevailed, which became impracticable due to increasing number of pupils and the time-consuming process of oral testing. Arithmetic, astronomy, geography, grammar, history and natural philosophy were the subjects of examination. This Boston examination project was highly commended by Horace Mann, the then Secretary of the Massachusetts Board of Education. He advocated (7) the replacement of oral by written examinations on various grounds like those of impartiality in scoring, thoroughness, no favouritism, availability of information to all, and amenability to scrutiny of questions by all.

Rev. George Fisher (1864) an English school master devised the first measurement device, the 'scale books' for measuring achievement in handwriting, spelling, mathematics, navigation, grammar, composition, drawing etc. Pupils' specimen were compared with standard specimen and ranking was done to determine numerical rating. Quantitative evaluation was done in terms of errors in performance as in spellings (8). Despite the great potential of Fishers' scale books, his ideas did not last long as he was ahead of his time in his educational thought.

E.E. White (1886), another American educational writer and school administrator wrote, "It may be stated as a general fact that school instruction and study are never much wider or better than the tests by which they are measured." He gave several special advantages of written tests such as tangible and reliable results, more accurate comparison of pupils' progress, revealing more clearly the defects of teaching etc.

Dr. J.M. Rice (1894) is considered the real inventor (9) of the comparative tests which became the basis of objective measurement. He administered a list of spellings to 13,000

students from 21 cities, spread over 16 months and in 1897 presented his paper to the Superintendence of the National Education Association. The findings that pupils who studied spellings for 30 minutes a day for 8 years are no better than those who studied for 15 minutes a day, for the same period, surprised every body. He was attacked and defamed for this 'heresy'. It was only a decade later that his objective method got attention in educational testing. The value of what we now call 'norm' is the resultant outcome of his investigation. Rice was not only a tester but a pioneer of progressive education.

During this century we may, therefore, witness three main developments. The first development is the popularity of the written examinations over the oral testing. The second development is the refinement of evaluative devices to ensure more objectivity in measurement. Thirdly, a breakthrough is made to compare performance of bigger groups of students, a step towards development of norms.

3. Developments during 20th Century

From the rudimentary approaches of the 19th century we now make a survey of the more comprehensive and scientific approaches that began at about the year 1900. Scates (10) traced the trends in measurement and evaluation from 1897 to 1946 into five periods of one decade each up to 1950. The last three decades are also added for this exposition to complete the era up to date.

3.1. *First decade* (1900-1910)

First decade witnessed the publication of the 'Futility of the Spelling Grind' by Rice (in fact 1897) which shattered the belief that the product of learning was intangible and only a teacher of a particular class could appraise it. The first Binet Intelligence Scale and the beginning of achievement tests in basic skills of Arithmetic and Language arts are the other special features of this decade which may be called an incubation period. It was in this period that the first book on Mental and Educational Measurement was written by Thorn-

dike (11) in 1904. This book influenced a lot the early development and popularisation of standardised educational tests. In 1908 Thorndike's student Stone (12) published the first Standardised Instrument in Arithmetic Reasoning while Thorndike (13) himself published in 1909 his Scale for the Handwriting of Children, which was the first Standardised Achievement Scale.

3.2. *Second decade* (1911-1920)

During this decade some of the intelligence tests and achievement tests were standardised. For the first time standardised achievement tests were used in the city of New York for a survey during 1911-13. Ayres published a handwriting scale in 1911, Hillegas a composition scale in 1912, Buckingham a spelling scale in 1913, Woody's arithmetic scale came in 1916 and Terman made famous the Stanford Revision by 1916. Introduction of group intelligence tests in World War I was another landmark of this period, besides the fight for objectivity in educational measurement. It was McCall (14) who was the first to express publicly in 1920 that teachers instead of using standardised achievement test can construct their own objective tests, then called the 'new type' of tests for classroom use.

3.3. *Third decade* (1921-1930)

In 1924 Ruch (15) published the book on informal objective tests with emphasis on teacher-made classroom tests. Test results were considered as only one type of evidence of pupils' achievement which reflected the tendency towards a broader scope of measurement than mere testing. Evaluation of all aspects of growth and the use of wider variety of tools and techniques characterised the informal objective testing movement of Tyler. This period, therefore, highlighted the broader concept of measurement so as to encompass a wider area of assessment of pupils' growth. More than 1,000 standardised tests had appeared by this time. Development of statistical techniques received a lot of attention during this period.

3.4. *Fourth decade* (1931-1940)

The contribution of Tyler (16) was felt more prominently in the field of objective testing although he made significant contribution to standardised testing also. He outlined the steps of construction and validation of tests by emphasising the instructional objectives as basis for both. Recognition of pupil behaviours as indicators of attainment of instructional objectives or desired learning outcomes was his outstanding contribution. The extension of the scope of achievement testing to the more intangible instructional outcomes was a precursor to the broad modern concept of educational evaluation. The Cooperative Test Services began publishing a number of parallel forms of tests for secondary schools. Personality tests like those of research, interest inventories, attitude scales, sociometric devices and anecdotal records were introduced during this period. By 1940, over 2,600 achievement tests were available in all traditional subject areas. During this period, the first Mental Measurement Year Book (1936) by O.K. Buros was published. In 1940 the National Council on Measurement in Education was established.

3.5. *Fifth decade* (1941-1950)

Whereas the last decade of the 1930's was a period of extension and expansion, the 1940's were marked by maturing and refinement of tools and techniques. Big programmes like the Appraisal of Activity Programme (17) for the Elementary Schools and the famed Eight Year Study (1942) of 30 high schools (18) involved a comprehensive design. Under the direction of Tyler a series of instruments measuring students' ability to reason, apply, interpret etc. were developed under the Eight Year Study. Under the Appraisal of Activity Programme, basic skills in reading, language arts and arithmetic, map reading, using indexes, interpreting charts, graphs and tables, interests, personal social adjustments, creative expression etc. were tested.

The pupil evaluation movement is exemplified by the work of Lindquist and Tyler which not only indicated the neglect of significant areas of pupils' behaviour but also led to the increasing emphasis on assessment of higher mental processes. In 1947,

the Education Testing Service was established. This concludes the five decades of developments classified by scales. We may extend it to the remaining three decades to make the account up-to-date.

3.6. *Sixth decade* (1951-1960)

This period marks the post-independence era. The publication of a Technical Standard for Achievement Tests, AREA-NCME, in 1955, Bloom's Taxonomy of Educational Objectives: Cognitive Domain (19) in 1956 and Project Talent (1959) to identify, develop and utilise human talents, were important landmarks in the history of educational measurement. The implications of Bloom's Taxonomy in measurement and evaluation will be dealt with at a greater length while dealing with educational objectives.

3.7. *Seventh decade* (1961-1970)

The outstanding contribution during this period is that of the publication, Taxonomy of Educational Objectives: Affective Domain (1964) by Krath-Whol (20). This publication gave a new dimension to evaluating students'on appreciations, interests, attitudes, values etc. Its details are also discussed in the chapter on objectives. It was during this period that the first issue of the Journal of Educational Measurement (1964) was published, providing, thereby a forum for discussion and exchange of researches in the field of measurement. A national assessment of educational progress also began during this period. A taxonomy in Psychomotor Domain was published in 1966 by Simpson (21) which provided the hierarchical classification of the goals of Psychomotor Domain.

3.8. *Eighth decade* (1970-1978)

Another taxonomy in Psychomotor domain devised by Harrow (22) in 1972 provided still another classification of psychomotor goals and a basis for measuring practical skills in the field of sciences and practical arts. Not much work has been done in this area in terms of hierarchical testing. During the seventies emphasis on identification and nurturing of talents continued to

attract more and more attention. An important inroad has been made by the accountability aspect in the field of evaluation studies concerned with the various curriculum projects. The emphasis is shifting from the psychometric approach of evaluation in schools, to that of the system analysis approach in which more stress is being laid on opinion of personnel who use various tools and techniques rather than on the tools and techniques they use. Testing is only one part of evaluation.

3.9. *Changing foci*

Achievement testing is only one aspect of evaluation. Intelligence testing, aptitude testing, personality testing are other areas which have their own history of development. In the preceding pages a brief historical retrospect has been given relating mainly to the birth, infancy, adolescence and adulthood of the testing movement. Since the present compendium is focussed on the examination reform movement in India, further details about mental measurement is not called for. Before taking up the case history of the evaluation movement in India we may like to identify the changing emphases or foci of measurement as under:

(i) Discrimination of persons on the basis of oral expression.
(ii) Selection or classification of persons on the basis of written ability.
(iii) Survival skills as a basis of testing, for physical development.
(iv) Written word in preference, to oral word for objective measurement.
(v) Comparative performance for developing group norms.
(vi) Empirical studies for standardisation of tests for wider use.
(vii) Cognizance of intangible and tangible instructional objectives for teacher-made tests.
(viii) Refinement of tools and techniques of testing
(ix) Use of statistical methods in educational measurement.
(x) Publication of tests, journals and tests.

Two major trends that one can observe from these foci are

the broadening concept of measurement and the increasing concern for making measurement more and more scientific.

4. Examination Reform in India

4.1. *Pre-independence developments*

(i) *The Wood's Despatch* (1854)

The Entrance examination started with the establishment of three universities in the presidencies of Bombay, Calcutta and Madras in 1857, the year of the Indian mutiny. This examination was as a sequel to the recommendation of the famous Wood's Despatch (23) of 1854. These 'Affiliating Type' universities at that time were founded to ascertain, by means of examination, the persons who had acquired proficiency in different branches of learning and 'rewarding them by academic degrees', as evidence of their respective attainments. The main objective of this Entrance examination was to select, from high schools, the students who were suitable for admission to higher education. Thus the first public examination was instituted to certify students for entry into the universities. The courses of study being only academic and unrelated to life, the instruction was mainly dominated by the Matriculation examination (Entrance) not only at the Secondary but even in Primary schools. As reported in the Review of Education in India (1886) the Matriculation examination led to six external examinations from Matriculation downward to the lower primary.

(ii) *Hunter Commission* (1882)

This commission (24) made a valuable survey of the secondary schools of that time and recommended two types of avenues, one leading to the entrance examination and the other of a more practical nature leading to non-literary vocational persuits. Secondary education on the basis of the grant-in-aid system was considered as an alternative to the growing cost of the expanding secondary education. But hardly any province adopted vocationalisation and education continued to be mostly academic, completely dominated by the Matriculation examinations.

(iii) *Indian University Commission* (1902)

This Commission (25) reviewed the position of the universities regarding the higher grades of examination. It pointed out that, "The greatest evil from which university education suffers in India is that teaching is subordinated to examination and not examination to teaching." In contrast to the tradition of general education in India, examinations started dominating the whole system of education to such an extent that they became instruments of general education. Assessment of grants to aided schools by results, further aggravated the demoralising influence of examinations on education. As reported in the Fifth Quinquennial Review (1902-1907) only two examinations, the Primary and Anglo-Vernacular were retained for the primary stage and the secondary stage, to meet partially the condemnation of the too many external examinations.

(iv) *Lord Curzon's Resolution* (1904)

It suggested (26) introduction of alternative courses to meet the needs of students who are destined for commercial or industrial pursuits. As a result, certain departments of education in Bombay, the Central Provinces and the United Provinces instituted many schemes of school-final examinations. But these examinations could not be popular as the matriculation examination continued to dominate as the scene and get preferential treatment for admission to the universities.

(v) *Resolution of 1913*

From 1854 to 1913 there was a phenomenal expansion of secondary education which resulted in the deterioration of the quality of output. This, according to the 1913 Resolution (27), affected the standard of higher education. This resolution appreciated the need of an external examination as a means of maintaining standards on the one hand and its limitation of neglecting the day-to-day progress of pupils on the other. The maintenance of records of the progress and conduct of pupils, the use of written as well as oral examination, the strengthening of supervisory staff and the need for making internal assessment a pre-requisite to present candidates for external examinations, were the major recommendations. The combination of external and internal assessment was considered as the

most important reform required in the secondary education, next only to the pay and prospects of teachers, according to this resolution.

(vi) *Calcutta University Commission* (1917-1919)
This commission popularly known as the Saddler Commission (28) opined that a mechanical system of marking encouraged memorisation and deteriorated mental development. The need for a small board of examination whose function should be to criticise and suggest improvement, was emphasised. It made the following important recommendations:

(a) That an institution of Intermediate colleges be created to provide instruction in arts, science, medicine, engineering, teaching etc. This stage may be considered as a dividing line between the university and the secondary courses. Admission to the universities should be made after passing of the Intermediate examination.
(b) A Board of Secondary and Intermediate Education consisting of representatives of Government, University, High schools and Intermediate colleges be established to administer and control the Secondary Education.

The Intermediate Education got dissociated from the University Education. Dacca, U.P., Panjab and Bihar universities gave this trial. U.P. created Board of High and Intermediate Education. Most of the other provinces did not take up and universities continued to control Intermediate examinations.

(vii) *Hartog Committee* (1929)
This Committee (29) reviewed the position of education in the country and suggested a diversified curricula. The Matriculation examination was found to still dominate the secondary courses. The Committee recommended the retention of more boys intended for rural pursuits in the middle vernacular schools. Diversification of more boys to the commercial and industrial careers by providing alternative courses was recommended in view of the high percentage of failures and rush to the university education.

(viii) *The Sapru Committee* (1934)
Appointed by the U.P. government, this Committee (30) observed the practice of preparing students for examinations and degrees rather than for avocations. It recommended the diversification of courses at the secondary stage; the abolishing of Intermediate colleges with the introduction of 11 year of schooling as 5+3+3 followed by 3 years of degree courses. The Abbot Wood Report (1936-37) recognised the problem of vocationalisation and suggested a complete hierarchy of vocational institutions parallel to the hierarchy of institutions of general education.

(ix) *The Sargent Report* (1944)
This was the post-war educational developments report (31) of the Central Advisory Board of Education. It visualised a system of universal, compulsory and free education for all boys and girls between the ages of 6-14, middle or senior basic school being the terminal stage for the majority. Academic and Technical courses were recommended in the 11 year schooling proposed in this report.

4.2. *Post-independence developments*

(i) *Provincial Committees* (1948)
The Acharya Narendra Dev Committee (1948) appointed by the U.P. government on reorganisation of primary and secondary education (31a) recommended the introduction of cumulative records, rules of promotions, use of written, oral and practical tests, use of intelligence and aptitude tests and the establishment of Bureaus of Examinations. At the same time in C.P. and Berar, a Committee (31b) on the reorganisation of secondary education recommended the following:

(a) Classification of purpose of examination in each subject and its mode.
(b) Maintenance of records to assess objectives, not testable through written examinations.
(c) Valuing of scripts by two evaluators and of moderation procedures.
(d) Careful selection of paper setters and examiners.

(e) Use of written, oral, practical examinations and records of pupils.
(f) Trial of symbolic system of marking except for the top few.

(ii) *University Education Commission* (1948-49)
This commission popularly known as the Radhakrishan Commission (32) recognised that examinations as they were, formed one of the worst features of Indian education because of their pernicious effects on the whole teaching-learning process. The need was felt to make examination relevant to the ends in view. They were so much alarmed with their domination that the view was expressed that "if we are to make a single recommendation for the whole university education it should be that of improvement of examination". The need for objective testing, appraisal of existing practices, appointment of full-fledged boards of examiners in each university, the construction of batteries of psychological and achievement tests, the delinking of university degrees from administrative services and credit to classwork, were the major recommendations made by this Commission. A scheme of self-contained units of work and allocation of one-third marks for course work for each subject in B.A./B.Sc., were also recommended.

(iii) *The Secondary Education Commission* (1952-53)
A diversified curriculum followed by a public examination at the end of 11 year schooling was the main viewpoint expressed. Other major recommendations of this Commission (33) were as under:

(a) Reducing number of external examinations and the subjectivity of essay type questions by introducing objective questions.
(b) Maintenance of systematic records about pupils' growth in all aspects during the whole year and credit to this internal assessment in the final evaluation.
(c) Use of symbolic rather than numerical marking in external as well as internal examinations.
(d) Indicating internal test data and other records along with the external examination results in the certificate awarded.

(e) Introduction of compartmental examinations at the final public examinations.

(iv) *U.P. Committee on Reorganisation of Secondary Education* (1953)

This committee (34) reflected on the invalidity, unreliability and subjectivity in assessment in the single examination of small number of questions marked by huge number of examinations. They recommended replacement of external examinations by teachers' assessments made during the session. The need for coordination of standards of assessment of different teachers in a subject from different schools and the use of scaling was suggested in this record.

(v) *Report of the International Team* (1954)

This team suggested abolition of private tuitions. Any number of examination subjects may be taken like the G.C.E. of England and it may be spread over stages. For development of objective type tests the establishment of an All India Educational Research Centre was suggested by the team.

(vi) *All India Council for Secondary Education* (1955)

With the establishment of A.I.C.S.E. its first meeting (35) was held on October 3-4, 1955 and it recommended the following:

(a) Secondary education boards should undertake research in this field by setting up a Research Unit.
(b) Training colleges should work in close collaboration with the Boards and conduct research, especially for improving essay type questions.
(c) A Committee should be set up to advise and report to the Council on the matter of examinations.

As a sequel to these recommendations, a seven-men Committee was set up. The terms of reference was as follows:

(a) Improvement of written examination and the feasibility of the objective type questions.
(b) Development of a scheme for improving practical examinations.

(c) Study evaluation examination criteria of States with a view to improving standards.
(d) To determine usefulness of school records for selection of students to secondary schools.
(e) To develop achievement tests by State Boards.

(vii) *Bhopal Seminar* (1956)

This seminar held in February 1956 is considered a landmark in the history of Examination Reform. It recommended the setting up of a section in A.I.C.S.E. comprising members from the A.I.C.S.E. and coopted experts to coordinate work of State Research Bureaus and to advise the Council in maintaining and improving standards. This section should be assisted by a Bureau of Examinations. The main function of this section envisaged by them was coordination, collection and dissemination of information about improved tests and research work of the state bureaus besides advising on improvement of standards. An Examination Committee appointed by the Chairman of A.I.C.S.E. met at Baroda in December 1956 and recommended the establishment of the Examination Unit. Major recommendations of the Bhopal Seminar were as follows:

(a) Inclusion of short answer and objective type questions besides the essay type in the external and internal examinations.
(b) Weightage of at least 20% to the sessional work based on cumulative records.
(c) Establishment of a Bureau of Examination Research by each Board of Secondary Education which should work in close collaboration with the training colleges, university departments of education and other state agencies engaged in research.
(d) Modification of the internal school examinations to ensure successful working external examinations.

(viii) *Visit of Dr. B.S. Bloom* (1957-58)

A.I.C.S.E. was fortunate to get the expert services of the world-renowned authority on examinations, Dr. B.S. Bloom, Head, Board of Examiners of the University of Chicago. He visited

schools in India and met teachers, teacher educators and administrators and conducted six evaluation workshops in which about 200 teachers participated. In these workshops he emphasised the integral relationship of the educational objectives, learning experiences and evaluation procedure, that formed the guiding philosophy of the programme of Examination Reform propounded by Dr. Bloom. Accordingly, a plan of action was formulated which formed the foundation of a countrywide movement of Examination reform. A ten-year phased programme was chalked out which included three main aspects:

(a) Setting up of realistic, significant purposes and goals of learning.
(b) In-service training of personnel to implement those purposes.
(c) Development of internal and external evaluation procedures to serve those purposes.

This phased programme envisaged that examinations could not be eliminated altogether (35a). It was visualised that changes in initial three to four years would be slow but were intended to bring a major reorientation and new attitudes. Evaluation procedures would always be somewhat in advance of the development of learning experiences. New type of questions would be introduced at first on a small scale and then would be gradually increased at the successive examinations. The detailed action plan was later on put before the first conference of the Chairmen and Secretaries for approval.

(ix) *First Conference of Chairmen and Secretaries* (1957)
This conference held at New Delhi in April 1957 (36) was the one where Dr. Bloom presented the full plan of action for scrutiny and acceptance. The following recommendations are worth mentioning in this connection:

(a) General approval of the programme of action obtained by Bloom.
(b) Endorsement of the view about evaluation as integral part of the total educational process.

(c) Setting up an Examination Unit at the A.I.C.S.E.
(d) Implementation by the state boards, the recommendations of the Bhopal Seminar relating to cumulative records, supply of achievement tests to schools, weightage to school assessment in external assessment and the use of objective type and short answer type questions in external examinations.
(e) Setting up of research units by the boards to undertake studies on examinations.
(f) Agreement on a 10-year phased programme of improvement in gradual stages.

(x) *Establishment of Pilot Examination Unit* (1958)
Consequent upon the recommendation of the first conference a Pilot Examination Unit was set up in A.I.C.S.E. which later on became a part of the Directorate of Extension Programme of Secondary Education, better known as D.E.P.S.E. Five evaluation officers recruited for the unit were posted to five zones of the country at the Extension Services Centres. During five months 60 orientation workshops were organised for 1500 teachers. Ten more officers of different subjects were sent to U.S.A. for training at Chicago under Dr. Bloom. On return they joined D.E.P.S.E. as Evaluation Officers. They were further groomed in India for one month by involving them in different seminars and workshops organised for the purpose. Two more regional workshops were held at Hyderabad and Chandigarh for selected teachers of secondary schools in eight subjects under the guidance of Dr. Bloom. It was assumed that external examinations in the country would continue and the use of objective based questions would bring educational improvement. The special role of teacher training colleges was emphasised in developing learning experiences, training of preservice teachers in evaluation and the need for inservice education of teachers in this field. The major functions of the Pilot Examination Unit were proposed as under:

(a) To work with teachers and training college lecturers in order to help them to identify, clarify and enlarge instructional objectives, develop learning experience and improve

evaluation tools.

(b) To work with Boards of Secondary Education to improve their paper setting and other examination practices.

(c) To work intensively with selected schools to try new practices in classroom.

(d) To train persons for State Evaluation Units.

(e) To develop item pools in various subjects.

(f) To develop tools for internal assessment.

(g) To conduct research related to internal and external examinations.

(h) To publish literature on educational evaluation,

(xi) *Second, Third and Fourth Conferences* (1958, 1959, 1961)

Among others, the following recommendations were made by these conferences (37).

(a) Establishment of State Evaluation Units.

(b) Development of a pool of test materials.

(c) Including the new concept of evaluation in B.Ed./B.T. syllabus as early as possible.

(d) Organise short-term courses in evaluation for trained teachers.

(e) D.E.P.S.E. should give publicity to disseminate the new evaluation concept.

(f) Setting up of a Central Bureau of Curriculum Development.

(g) Two-level examinations in English and Mathematics.

(h) Examine possibility of announcing performance of students rather than declare them as pass or fail.

(i) Training of paper setters and examiners.

(xii) *NCERT and Examination Reform* (1961)

A number of departments of the Ministry of Education like the National Institute of Basic Education, National Institute of Audiovisual Education, D.E.P.S.E., Central Bureau of Educational and Vocational Guidance, Central Bureau of Textbooks, Curriculum and Research etc. were merged and some new departments were created under the umbrella of the National Council of Educational Research and Training (N.C.E.R.T.) in

September 1961. In another reorganisation, an independent unit, the Examination Reform Unit, came into being. Soon after, it became a part and parcel of the newly created bigger department named as Department of Curriculum and Evaluation. Even this department was later on abolished on the recommendation of the Nag Chaudhari Committee (38) appointed to review the work of the N.C.E.R.T. However, the new Department of Textbooks created in the N.C.E.R.T., continued work in the field of Examination Reform but with a much low zeal. In a yet another restructuring, the Department of Textbooks was reorganised with the setting up of two new parallel units pinched off from the parent Textbook Department and were renamed as the Examination Reform Unit and the Examination Research Unit. These units were again integrated into one department of Measurement and Evaluation including the sister unit of National Talent Search. With the last reorganisation of the N.C.E.R.T., Measurement, Evaluation, Talent Search unit and Survey unit were merged in 1984 to form a bigger Department of Measurement, Evaluation, Survey and Data Processing.

(xiii) *5th, 6th, 7th and 8th Conferences* (1963-68)

Among others, following were the major recommendations of these conferences (39).

Fifth (*Delhi—1963*)

(a) To set up a seven-member Standing Committee to expedite Examination Reform Programme.
(b) Each board should prepare a concrete plan of action to introduce reform in the existing system of examinations.
(c) The Central Examination Unit to serve a clearing house function.

Sixth (*Poona—1964*)

(a) To take steps to improve practical examinations in science subjects and oral examinations in languages.
(b) The boards should provide funds for training of paper setters, examiners and moderators.
(c) The staff of the State Evaluation Units should be augmented and funds be prov ided for their effective working.

Seventh (*Mysore—1965*)

(a) To train pupil teachers of B.Ed./B.T. classes in improved evaluation procedures.

(b) To prepare and circulate the paper on scaling procedures by the Central Examination Unit and a brochure on development and use of a library of test material.

(c) The Central Examination Unit to coordinate the research activities of State Evaluation Units and the Boards of Secondary Education.

Eighth (*Ajmer—1967*)

(a) State Evaluation Units be attached to the Boards of Secondary Education.

(b) Syllabi in different subjects should include specific instructional objectives and detailed content outline.

(c) Disparity of using cut scores by different boards to determine pass or fail in a subject should be removed.

(d) A State Evaluation Organisation as recommended by the Education Commission for each state, may not be created but the functions of State Evaluation Units be transferred to State Boards of School Education.

(xiv) *Education Commission* (1964-66)

This commission (40) considered evaluation as a continuous process integral to the total system of education which exercised a great influence on the study habits of students, instructions and mode of evaluation. The following recommendations are relevant to this paper:

(a) To improve written examinations for making them valid and reliable measure of educational achievement.

(b) To devise tools and techniques for those area of pupils' growth which cannot be measured by written examinations.

(c) At the lower primary stage, evaluation should help students to improve their achievement in basic skills and in the development of right habits. At the higher primary stage, use of oral tests, diagnostic tests and use of cumulative record be made.

(d) There should not be external examinations at the end of

S.C.E.R.T., West Bengal Library

the primary stage. However, to maintain uniformity of standards, periodical surveys may be conducted by district school authorities using refined tests prepared by the State Evaluation Organisation.

(e) For inter-school comparability, a common external examination may be held by district authorities using standardised or refined tests. This is besides the merit scholarship examinations.

(f) A certificate be issued by the Board to indicate students' performance in different subjects without any reference to pass or fail. A student may be allowed to reappear in all or in one or more subjects to improve his performance.

(g) Schools should issue a separate certificate showing the record of internal assessment as contained in a cumulative record.

(h) A few selected experimental schools be established and given freedom to prescribe their own curriculum, textbooks and assess their own students and award certificates at the end of class X which may be considered equivalent to the Board certificate. However, the State Board will issue certificate on the recommendation of the schools.

(i) Comprehensive internal assessment of all aspects of students' growth should be undertaken and this should be descriptive as well as quantitative.

(xv) *National Policy on Education* (1968)

A parliament committee was set up to examine the Education Commission Report. According to the national policy on education developed by this committee, it was emphasised that, "A major goal of examination reform should be to improve the reliability and validity of examination and to make evaluation a continuous process aimed at helping the student to improve his level of achievement rather than certifying the quality of his performance at a given moment of time" (41).

(xvi) *V.K.R.V. Rao Committee on Examinations* (1971)

With the increase in malpractices and incidents of violence in external examinations, the Central Authority Advisory Board at its 35th meeting expressed its concern and requested its

S.C.E.R.T., West Bengal
Date 29.1.91

Chairman to appoint a committee on examinations. A committee of seven members was appointed with Union Education Minister as Chairman. The other members were the Education Minister of Andhra Pradesh, (Vice Chairman); the Education Ministers of Assam and Bihar; the Chief Executive Councillor, Delhi, Mr. A.E.T. Barrow and Prof. S.V.C. Aiya, Director, NCERT as Member-Secretary. The committee had six meetings between August, 1970 and June, 1971. The major recommendations of this committee (42) were as under:

(a) To enact legislation empowering examining authorities to take suitable measures regarding prohibition of weapons, assembly of persons, protection of invigilators etc.
(b) To institute more than one examination where the number exceeds 10,000 school students or 1,000 college students.
(c) Precautions for setting papers be taken well in advance by requiring model answers for them. Teachers should be free to comment on the paper after it is administered in the examination.
(d) Method of spot evaluation at a central place should be adopted and results should be declared subject-wise and in the form of grades.
(e) The certificate should indicate separately, in two columns, the results of public examination and that of internal assessment.
(f) There should not be too many public examinations except one at the end of middle/upper primary; another at the end of the secondary and the third at the first degree college. All other assessments should be only internal.
(g) Admission to colleges, including professional ones, and recruitment to services should be on the basis of selection tests.
(h) Both Central and State Governments should earmark funds separately for research on examinations which should be undertaken in a coordinated manner with the Boards and Universities.
(i) A novel idea of organising and conducting examinations is suggested. This includes multiple paper setting and their use in place of one paper, involving practising teachers in

paper setting by inviting questions from them with model answers and their scrutiny by a board to finalise papers for public examinations; decentralisation of conduct of examination when there are more students and use of project, work for assessing creativity, initiative and special skills.

(xvii) *University Examination Reform Action Plan* (1972)
A working group set up by the Ministry of Education and Social Welfare brought out the document, 'Examination Reform — Plan of Action' (43), and it was endorsed by the Commission at its meeting held on August 2, 1971. In 1974 the U.G.C. convened four zonal workshops relating to examination reform, autonomous colleges and postgraduate education, at Chandigarh, Madurai, Ahmedabad and Bhubaneshwar. A number of suggestions were made relating to internal assessment, question banks and grading system. Preliminary try out of internal assessment in unitary universities, the immediate feedback of results, opportunity to students to represent against the awarded grades, the maintenance of record for scrutiny, keeping internal and external evaluation marks separate, the development of question banks at each university and their effective use in paper setting, a seven point scale for grading, the mechanics of grading and their interpretation and setting up of a grievance machinery to redress students grievances, were the major recommendations.

(xviii) *Approach Paper on Ten-Year School* (1975)
The Ministry of Education and Social Welfare constituted an expert group in 1973 to develop the curriculum for the 10+2 pattern of education. This group was further expanded in 1974 by including experts from the N.C.E.R.T. who drafted a version of the curriculum in 1972, revised in 1973. The final document called the Curriculum for the Ten-Year School—A Framework, was published in (1975) by the N.C.E.R.T. (44). This document is characterised by the compositness of its approach, flexibility within the framework of acceptable principles and values, the relevance of the disciplines of the curriculum, the unit approach to teaching and learning and using evaluation as feedback. The major suggestions relating to examination reform were as under:

(a) Evaluation should give reliable evidence on attainment of all instructional objectives by the use of a variety of tools and techniques.

(b) Evaluation by teachers should be done continuously and the results feedback to pupils.

(c) Passing all subjects at a time should not be insisted upon and a semester system may be adopted.

(d) At the primary stage evaluation should be integrated with the process of learning based on a continuous system of recording pupils' progress using observation and oral tests; written and practical tests may be used at the middle and secondary stage in addition to other tools and techniques.

(e) There should be no 'pass' or 'fail' and students be graded on a five-point rating scale. Students may be allowed to improve their grades.

(f) A subject-wise/unit-wise record of students' performance in scholastic and non-scholastic areas should be kept regularly as part of the internal assessment.

(g) The system of internal assessment should be strengthened and measured to reduce biases be taken till a stage comes when external examinations become redundant and are abolished.

(h) District Education Officers/Inspectors may set up a committee to make sample checking of question papers and answer scripts to check biases and ensure proper evaluation. At the school-complex level and state level also, such committees could be established.

(i) Community meetings may be held from time to time to acquaint them about the mode of evaluation adopted and its use in improving pupils' achievement.

(xix) *Approach Paper on Plus-Two Stage* (1976)

Following on the heels of the 1975 Approach Paper on the Ten-Year School, another approach paper was published by the N.C.E R.T. on Higher Secondary Education and its Vocationalisation (45) in 1976. This paper though it dealt mainly with the pattern of academic and vocational courses at the plus-two stage and other problems relating to vocationalisation, did make some recommendations on evaluation of the courses. These were

as under:

(a) There is need for dispensing with the public examination at the higher secondary level by adopting a continuous system of evaluation with a provision for checks and balances. Effective supervision would be needed to maintain a high standard of performance.
(b) A flexible system of two to six semesters be introduced and students' performance on all these semesters be recorded on a result-card using a seven-point scale for grading.
(c) There is need for establishing equivalence among vocational diplomas and certificates issued by various agencies. The National Council of Vocational Education should take up this work.
(d) Since a student is expected to reach a particular level of competence to qualify for a certificate, students may be required to accumulate credit points for the course which is divided into various units or credit courses, requiring a different duration of time. Cumulative credit may be worked out subject-wise and a rule for getting a certain number of credits as a minimum requirement for passing the +2 stage may be framed.

(xx) *National Policy on Education* (1986)
In the N.P.E. evaluation process and examination reform envisages the following emphases:

1. Assessment of performance is an integral part of any process of learning and teaching. As a part of sound educational strategy, examinations should be employed to bring about qualitative improvements in education.
2. The objective will be to recast the examination system so as to ensure a method of assessment that is valid and reliable measure of student development and a powerful instrument for improving teaching and learning. In function terms it would mean:

   2.1. Elimination of excessive element of chance and subjectivity.

2.2. The de-emphasis of memorisation,
2.3. Continuous and comprehensive evaluation that incorporates both scholastic and non-scholastic aspects of education spread over the total span of instructional time.
2.4. Effective use of evaluation process by teachers, students and parents.
2.5. Improvement in the conduct of examinations.
2.6. The introduction of concomitant changes in instructional materials and methodology.
2.7. Introduction of semester system from the secondary stage in a phased manner.
2.8. The use of grades in place of marks.

3. The above goals are relevant both for external examinations and evaluation within educational institutions. Evaluation at the institutional level will be streamlined and the predominance of external examination reduced.

5. The Changing Emphases

The last one and a quarter century of the history of examinations in India indicates a number of trends. Shifts in the emphases is indeed due to visualisation of a broader view of education and its purposes on the one hand and the influence of examinations of the total educational process on the other. Education is no longer equated with rote memorisation and examinations are no longer considered as the end of teaching-learning. We may trace the following changes in the emphases during this period:

(a) The concept of the traditional examination which dealt only with collecting the information about pupils' achievement is giving way to a more comprehensive concept of evaluation which deals with forming judgements and decision-making besides the information-gathering process.
(b) There is extension in the scope of assessment of pupil growth from mere scholastic achievements to include non-scholastic traits.
(c) The purpose of examinations is no longer limited to

measurement of students' learning but is extended to improving their learning.

(d) The function of examination is not only to test for factual information but encompass testing of all the instructional objectives.

(e) Instead of treating evaluation as an end of the course activity, it is being considered as an integral part of the teaching-learning process thereby ensuring a continuous process of students' appraisal.

(f) Besides written examination other tools and techniques like oral testing, practical tests, observations etc. are now increasingly considered indispensable for students' evaluation.

(g) More attention is being paid to refining the measuring instruments to make them valid and reliable tools of evaluation.

(h) There is a tendency to depend more and more on teachers' evaluations with a focus on total internal assessment.

(i) Increasing need is being felt to discourage irreversible judgements of failure and make provision for students to improve their performance.

(j) There is a clear shift in the analysis of test results for diagnosing pupils' difficulties in place of mere grading of their achievement.

(k) There is a trend, especially in the measurement of non-scholastic traits, to use the five-point or seven-point grading system in place of numerical marking which is responsible for misclassification of students.

(l) A clear emphasis is observable on the feedback of results for reinforcement, improvement of students' learning and the teaching-learning process.

(m) Meaningful certification of students' achievements is the visible trend. This is being sought for by reflecting a more comprehensive and multifacet growth of students as revealed by external evaluation and internal evaluation separately.

A chronological summary of significant events in the history of examination reform is given in the annexure that follows.

It is thus obvious from the historical development of measu-

rement and evaluation that the need for integrating evaluation with the educational process is being realised more and more. This means that it is the class teacher alone and not the external evaluator who can achieve this goal. He has to be made the main arbitrator of a student's appraisal. In this task, the role of his colleauges, students, supervisors and administrators has to be appreciated. From examination to evaluation is a long distance to be traversed. A number of new ideas and concepts are involved. Unless the new terminology that is used in current literature is understood, it is difficult to appreciate the evaluative process. Therefore, a discussion on the basic pedagogical concepts in evaluation is necessary before explaining the process of evaluation.

ANNEXURE

Chronology of Education Events

Year	Significant Recommendations	Recommending agency
(1)	(2)	(3)
1854	Institution of the Entrance Examination as a first public examination.	Wood's Despatch
1882	Suggestion for practical and academic courses. Domination of Matriculation Examination.	Hunter Commission
1902	Assessment of grants to scools, by results.	Indian University Commission
1904	Many new schemes of examinations alternative to Matriculation.	Lord Curzon's Resolution
1913	Combination of external and internal assessment.	Resolution of 1913
1919	Creation of Intermediate colleges to act as basis for admission to university.	Calcutta University Commission (Sadler).
1920	To reduce incidence of failure and limit rush to universities by provision of alternative courses.	Hartog Committee
1934	Abolishion of Intermediate colleges and case for 11 year (5+3+3) schooling.	Sapru Committee
1937	Hierarchy of vocational institutions parallel to hierarchy in general Education.	The Abbot Wood
1944	Universal, compulsory and free education for age group of 6-14.	Sargeant Report
1948	Improvement of examinations by appointment of full-fledged board of examiners in each university and allocation of one-third marks for classwork.	Univesity Education Commission
1952	Introduction of internal assessment data in the certificate along with external data and a case for use of symbolic grading.	Secondary Education Commission
1954	Setting up of All India Council of Secondary Education.	Ministry of Education
1956	Setting up of a section in A.I.C.S.E. to promote research work and examination reform programmes.	Bhopal Seminar
1957	Dr. Bloom's visit and development of ten-year phased programme of examination reform and its approval by the first con-	

(1	(2)	(3)
	ference of Chairman and Secretaries of Secondary Education.	
1956	Establishment of Pilot Examination Unit in A.I.C.S.E.	
1961	Establishment of the National Council of Educational Research and Training. Establishment of the State Evaluation Units.	Ministry of Education
1963	Setting up a Standing Committee on Examination Reform Programme.	
1964	Rajasthan Board of Secondary Education launches Examination Reform Project in collaboration with N.C.E.R.T.	N.C.E.R.T. and Rajasthan Board of chool Education
1966	Comprehensive evaluation of students by using variety of tools and techniques in internal assessment. Extermination of external examinations from the primary stage and use of periodic surveys or common external examination for comparability and maintenance of standards.	Indian Education Commission
1968	National Policy on Education emphasising the improvement of quality of instruments and using evaluation for improvement of students' achievement.	Parliament Committee
1971	Review Committee recommends multiple-paper setting, spot evaluation of scripts and separate indications of internal assessment and public examination performance in the certificate.	V.K.R.V. Rao Committee
1974	Development of action plan for Examination Reform at university level.	University Grant Commission
1975	Evaluation as feedback service and strengthening of internal assessment programmes for ultimate abolishing of external examinations.	N.C.E.R.T. Curriculum Framework for Ten-Year School
1976	Unit-wise flexible system of two to six semesters with seven-point grading procedure leading to cumulative credits forming the basis of promotions.	Plus-Two Approach Paper, NCERT
1978	Development of action plan for improving evaluation at the elementary stage.	Examination Reform Unit N.C.E.R.T.
1986	Integrated evaluation with teaching through continuous, comprehensive assessment in schools—Scaling of external examination marks, etc.	National Policy on Education

CHAPTER II

CONTEMPORARY CONCEPTS IN EDUCATIONAL EVALUATION

1. The Terminology Web

In the field of educational evaluation India is one of the developed countries. The work done during the last two decades bears testimony to this statement. Some of the neighbouring countries like Sri Lanka, Nepal, Malaysia and Afghanistan, have already received a lot of technical help in this field and got their top planners in education trained by us through the N.C.E.R.T. in the concept and techniques of educational evaluation. In recent years, especially during the present decade more and more concern for evaluation is being shown. The reasons may be the shrinking resources, dissatisfaction with educational dividends, decentralisation in development and implementation of educational programmes besides some other. Accountability has become of far-cry of most of the new projects of curriculum and in-service education programmes. With the widening concept of education, variation in the context, resources, process, competence of personnel and availability of data gathering devices, a number of new mothodologies of evaluation are being used. This has led to the emergence of many new concepts in educational evaluation. In fact a jungle of terminology has grown which has made it difficult for teachers to comprehend and differentiate between some of the current terms in use. It causes confusion sometimes even to the educational administrators and teacher educators. It is, therefore, desirable to identify the major

pedagogical concepts relating to educational evaluation, define them and establish relationships among them.

The terms in vogue are test, examination, assessment, appraisal, measurement, evaluation, diagnostic, formative and summative; formal, non-formal and informal; internal, external and exo internal evaluation; rational; experimental and illuminative; norm referenced, criterion referenced and self-referenced; process, product and collateral; objective based, goal free and pay-off evaluation; qualitative, quantitative and responsive; micro, macro and mega evaluation; cognitive, affective and psychomotor; teacher based, standardised and programmed tests, self evaluation, participatory and teacher dominated; pupil, curriculum and programme evaluation. Let us take up related concepts one by one and understand their connotation, scope and implications for the educational process.

2. Examination, Measurement and Evaluation

These are the most commonly used concepts and are often used interchangeably in educational parlance. The word 'examination' has often been used as a tool or instrument rather than as a process. Examination indeed denotes a process of collecting evidence about pupils' growth. It is a technique or a data-gathering device. As such it involves planning, construction of tools, administration, scoring of scripts etc. In this process one or more tools like a test or a question paper can be used besides teachers' observation or even students' product or performance, to collect the needed information. Take for example an examination in practical biology. When we ask our students to *examine* the given specimen and identify, it involves a process of observing, using a microscope or lense, understanding of relationships among different parts, recalling characteristics of a class of animals or plants. So unless this information is got, whether through the naked eye, a lense, a microscope or by previous knowledge, one cannot identify the given specimen. An examination is thus a process of gathering information about pupils' achievement and not a tool of measurement.

Measurement is an act of ascertaining of pupils' growth or

achievement. It deals with the comparison of quality with an appropriate scale for the purpose of determining the numerical value on the scale that corresponds to the quantity to be measured. According to Lindquist (1), "measurement is an act of assigning numerals to an object or an attribute according to certain rules." The numerals may be 1, 2, 3, 4, 5 or A, B, C, D, E or any other used for the purpose. Object or attribute to be measured may be achievement, physical efficiency, intelligence, aptitude or any personal or social quality. Rules of assignment of numerals may be for example to give A to only 7% of the students or to award full marks to each correct spelling. Thus measurement may be quantitative or qualitative in nature. Sometimes the word 'measurement' and non-measurement are used in place of the above two terms. In any case this is a concept which deals with the 'how much' of a thing. To determine this, one has to use some measuring device or tool which facilitates the collection of the needed information or to refine the act of measurement. A thermometer, a lens, a scale, or a question paper are all tools of instruments of evaluation which help us to make measurement or bring precision to it.

Measurement is, therefore, a broader term than examination which is only one of the many other techniques or processes of collecting information like observation, interview etc., which are involved in the act of measurement. Likewise, an examination is a broader concept than a tool like, say, a question paper which is one of the many tools concerned with the gathering of information.

Evaluation is a still broader concept than measurement. According to Popham (2) systematic evaluation consists of formal assessment of the worth of an educational phenomena. It denotes formal attempt to collect information and to reach a judgement about assessment of a merit. Tenbrinks (3) defines "Evaluation as a process of obtaining information and using it to form judgements which in turn are to be used in decision making." This term is used in close conjunction with measurement but it cannot be used interchangeably. Evaluation refers to the adequacy or worthwhileness of a performance. It involves value judgements which are based on measurement. The process

of determining the degree of adequacy depends heavily on the information obtained by formal measurement. Evaluation, therefore, involves value judgements which in turn are formed on the basis of measurement (quantitative) or non-measurement (qualitative). Measurement is, therefore, a pre-requisite for any evaluation. There cannot be any evaluation without measurement. However measurement is not synonymous with evaluation. Whereas measurement deals with the 'how much' of a certain thing, evaluation goes a step further to tell us of 'what value' is that measurement? The relationship between the two is depicted clearly in a mathematical fashion by Gronlund (4) as under:

$$\text{Evaluation} = \text{Measurement} + \text{Value judgement}$$

Let us take an example to concretise our understanding of the two terms. Suppose Anjum got 56 marks in a class test in English. The average performance of the class is 42. The highest score is 60 and the lowest score 15. A score of 56 is an index of her measured performance or her status. It does not indicate how good or how bad this score is. We have only measured and not found out the worth of that measurement. But on analysis when we find that she has secured the third position in the class and has a good performance with respect to the class, we are evaluating her performance, a value is being attached to it. If measurement is status determination, evaluation is worth determination.

From the above explanation we can identify the relationship between examination, measurement and evaluation. Evaluation is a three-stage process. Information gathering through examination or any other technique is the first stage. This information provides a score or letters or some sort of description which is the result of measurement (or non-measurement). This measurement provides the basis for forming judgements that reflects the worthwhileness or otherwise of the attribute measured.

3. Test, Assessment and Appraisal

According to Goods' dictionary, a *test* is a device or procedure for determining the truth, falsity, meaning of a hypothesis (—). Although the teachers are not aware of the type of hypothesis,

they generally try to test whenever a test is administered. Indeed they do test one or more of the hypotheses such as those relating to the development of the concepts, the attainability of the instructional objectives or the comparative achievement of two or more groups. A test is one tool of examination to collect a particular type of information about pupils' growth. Other devices of examinations may be observation, inquiry, analysis or opinions. A test could be oral, written or practical depending upon the nature of hypotheses to be tested.

Assessment is referred to as a teachers' estimate of a value against a certain standard. It is, therefore, only an estimate, not an accurate observation of an attribute. This term might have intruded into the educational parlance from the field of taxation where only a rough estimate is made about the property in question. This term suits very well in the context of internal assessment where effective assessment of non-scholastic traits is generally approximate and cannot be measured accurately like cognitive abilities.

Appraisal is another term which is used interchangeably with assessment and evaluation. It is generally defined as a formal and accurate valuation of some characteristics against some criteria usually made by persons familiar with such values (—). In contrast to estimation which can be informal and an approximation, appraisal connotes well defined criteria against which one is to estimate the worth of an attribute or a characteristic with a view to sell the idea, thing or material. In a way it is a more refined way of estimating the value of a thing by using such criteria as are supposedly developed by knowledgeable people.

4. Written, Oral and Practical Examinations

The nature and scope of these three types of examination is determined by the mode of collecting information about pupils' growth. If pupils' responses could be gathered through the written word then we call these as *written examinations.* If the medium of students' response is oral or aural then we term it as an *oral*

examination. If the medium of testing is the performance of the students then *practical examination* is the type. Written examination is normally undertaken with the help of paper and pencil. It may take the form of a unit test, a question paper or a written assignment of any kind. Since the written word is not necessary in case of such achievement tests where only diagrams are used as a medium, the use of the term, "paper-pencil," test seems to be more appropriate than the written examination. Most class tests and public examinations belong to this category of examinations.

Certain instructional objectives are difficult to assess through the written word because we are interested in the measurement of skills like those of speaking, listening, reading and oral communication. Here the medium of response is the oral word. This may take the form of conversation, dialogue, oral question or viva-voce in case of science subjects. Things which cannot be tested through the written word can be tested through oral means. That way oral examinations are more comprehensive in scope than written ones. But for the diagrammatic skills and written expressions all other objectives can be tested through these examinations.

Apart from the written or the oral word there is the medium of actual performance of the individual which forms the basis of examinations. Their scope is however limited to the assessment of psychomotor skills like those of observation, manipulation, dissection, collecttion etc. A practical examination indeed is the most reliable method of collecting evidences about pupils' growth as far as development of experimental skills are concerned. Their scope is limited to mainly science subjects, physical education, socially useful productive work etc. A well-thought out practical examination is the most comprehensive in that it not only involves testing of psychomotor skills but also some oral work (viva-voce) as well as some written work in the form of observational record and drawing work. A separate chapter is devoted to the practical examinations for discussing further details.

5. Cognitive, Affective and Psychomotor Outcomes Evaluation

Since Blooms' taxonomy of educational objectives (5) these words

are more generously associated with teaching and testing. *Cognitive evaluation* refers to the testing of those objectives which fall under the category of intellectual development e.g. knowledge, comprehension, application, analysis, synthesis and evaluation. These are the six objectives classified as cognitive objectives. Thus testing for these objectives has come to be recognised as cognitive assessment or evaluation.

Similarly, the assessment in terms of *affective* objectives has been recognised as affective evaluation. It covers those objectives which are concerned with the development of attitudes, interests, appreciations and personal and social development. It is this area of pupils' development which can be assessed better only through internal assessment. However, these are techniques too technical to be used by teachers in the classroom. *Psychomotor* assessment relates to those objectives which require students to perform some activity or do some practical work. This involves assessment of the product of performance as well as process of performance. Assessment of such area of development which involves motor activities or manipulation of objects or performance of a task like that of a physical activity, production of a model, handling of a microscope etc. involves on-the-spot observation and assessment of performance of students.

6. External, Internal and Exo-internal Evaluation

These terms are quite familiar to all teachers. However, there is definitely a confusion about their scope. At present external examinations are equated with only public examinations which become the basis of certification of students. All other examinations which come under the purview of the school are deemed as internal examinations. This is only one notion which is based on the agency conducting the examinations. From this it follows that an examination in a school, whether conducted by an outside person, or a teacher who does not even teach that particular class or a section, are all internal assessments. This view does not seem to hold ground because in all these cases the examiner is not connected with the instructional process of students who are being evaluated. The criterion indeed is the

knowledge on the part of the evaluator, of what has been taught and how it has been taught. It is only the class teacher who teaches the subject is aware of this fact. If an examiner is unaware of the unit objectives set in advance as also the learning experiences provided to the students, he is not in a position to evaluate them properly. Therefore, *internal evaluation* is one in which the teacher who teaches the students also evaluates them. All other types of evaluations, whether done in public or in schools, inside or outside the school, by a teacher of the same school or from a different school, are all *external evaluations*. Thus the three criteria for an internal evaluation are; familiarity with the teaching-learning activities of the class, setting of the question paper by the class teacher teaching that class and conduct of the examination vis-a-vis evaluation of scripts by the teacher himself as in the case of unit testing.

Exo-internal evaluation envisages a combination of both external and internal assessments. This combination can be in terms of the setting of the examination, conduct of the examination, evaluation of the scripts or for certification of students. The following are some of the possible practices in this regard.

(a) Papers are set externally but evaluation of scripts is done internally by the schools as is done by the Central Board of Secondary Education in subsidiary subjects like Sanskrit and Mathematics at the Higher Secondary stage.
(b) Papers are set externally, examination is conducted externally but the scripts are evaluated internally as done in some of the state examinations held at the end of the primary stage.
(c) Papers are set internally by schools by teachers not teaching the class but scripts are evaluated by the class teachers as is the case in many of the annual examinations upto classes IX.
(d) Papers are set internally by the class teachers but evaluation of scripts is done by the teachers not teaching that class.
(e) A team of class teachers sets the paper but evaluation is done either by class teacher or those who are not involved

in teaching.

(f) The paper is set by the class teacher of one section and is evaluated by class teacher of the other section.

There can be some more variations of this. As regards certification, there are again different patterns. One is the certificate issued by the Board which contains only the performance of students in the external examination. Two; along with external assessment there is also mention of the internal assessment performance in the same certificate. Three; the Board issues certificates on the external performance while the school issues certificates showing performance on the internal assessment. Four; certificates are only given by schools on the basis of their own examinations conducted during the session and students are promoted or detained on the basis of their own certification but the certificate is issued under the seal of the Board. This is the case in the two autonomous schools of Rajasthan located at Udaipur, Vidya Bhawan and Banasthali Vidyapith.

7. Formal, Non-formal and Informal Evaluation

Formal evaluation is the conventional, ceremonious form of evaluation having an established mode of its conduct. Rigid, punctilious and set rules are observed in carrying out formal evaluation. The formality may be in timing, frequency, mode of data collection, forming judgements or taking decisions. A public examination, a term test or a unit test are all part of the formal evaluation as also a standardised test. But the extent of formality in terms of rigidity, preciseness or explicitness pertaining to timing and mode of conduct varies in all these tests. A standardised test is more formalised than a public examination which in turn is more formal than a term test or a unit test.

Formal evaluation presupposes a planned and highly structured process of evaluation. Its place in the system is institutionalised and represents the official mode in vogue. This form is applicable to both curricular and co-curricular areas of pupil evaluation. Its place of conduct and time is predetermined. Tools of evaluation are specified before hand. Data are collect-

ed in a set form. Judgements are made in accordance with the established norms or tradition. The focus of such evaluation is on grading, classification and certification of students. Formal evaluation, therefore, refers to that mode of evaluation in which the various steps involved in the evaluative process are conventionalised and have acquired well defined and ritualised methodologies.

Non-formal evaluation seems to have evolved more recently in relation to non-formal education. Just as non-formal education has a flexible frame of purposes, curricula, processes, instruction and evaluation, the term non-formal evaluation may be viewed in that very context. In contrast to formal evaluation, non-formal evaluation is neither institutionalised nor that well structured. It may be planned or unplanned but it is not ceremonious and rigid in its operation. Its scope is not clearly delineated in advance but is adjustable to the situation and is thus need-based. Unlike formal tests, the non-formal tests are timed according to the need and progress of the subgroups or individuals. Spacing of tests is not determinable in advance. For data collecting we have to depend on sources other than the teacher also. Judgement-making rests more on individual performance rather than class performance. Thus self-referenced judgements are more valuable than class-referenced judgements. The purpose is mainly the improvement of individual's proficiency by regular feedback. This term may preferably be applied in case of such individuals or groups as are being educated through methods other than formalised schooling e.g. adult classes for functional literacy, correspondence courses, remedial classes etc. Non-formal evaluation, therefore, is unconventional, flexible, need-based and unceremonious in its purpose, scope, procedures and methodology of judgement-making. Its place in the traditional schooling seems to be out of tune and is difficult to justify unless close linkage between in the two systems is established.

Informal evaluation is as old as the reasoning power of man. Selection of appropriate jungle leaves for food, choice of a subject in school, rejecting a certain variety of fruit when compared to another, positive comment on a movie, making a

choice among different pieces of cloth for your garments, reflecting on Rita as a poor reader, saying, 'it is not a good buy', are all examples of an informal evaluation. We encounter such examples in our daily life from morning till evening. In all such evaluations one is neither conscious of the objective of evaluation nor is it a planned or deliberate effort to evaluate. It is all incidental and totally unstructured. Its scope is unlimited and cannot be circumscribed or determined in advance. The place of evaluation which is fixed in both formal and non-formal evaluation, is indeterminable in informal evaluation. Its timing cannot be settled or regulated. The mode of data collection rests mainly on one's observation or previous experience. Therefore, judgement-making is generally experience-referenced. The focus of informal evaluation is either on choosing among the alternatives or on the reflective judgements of the observer.

In a school there are a number of situations where informal evaluations are made by teachers about the students, fellow teachers and administration. Informal testing goes on in teaching through the use of oral questions. An unannounced testing of students on a particular topic or a concept, assessment of unaccountable home or school assignments, correcting students' pronunciation or intonation during reading aloud, rectifying students' mistakes in composition or in the practical class and scores of such other situations are instances of informal evaluation. In fact informal evaluation is a part and parcel of a teacher's personality, who tends to evaluate inside the class, outside it, in the home, in the market. It is thus an unintended, untimed, unplanned and unstructured form of evaluation. If the frequency of class-room informal evaluations is increased and the teacher becomes conscious of the role of such evaluation, students can be relieved, to a great extent, of the anxiety, fear, and phobia of formal examinations. If the formal examinations are made more and more informal and conducted in a non-formal manner, teachers and not the students will be more afraid of preparing for such examinations.

8. Summative, Formative and Diagnostic Evaluation

The forms 'summative' and 'formative evaluation' were for the

first time conceptualised by Michael Scriven (6) in his classic (1967) essay on the methodology of evaluation. According to him, summative evaluation refers to the assessment or worthwhileness of the instructional programme which is completed while formative evaluation refers to the assessment or worth of the instructional programme which can still be modified. A summative evaluator gathers information and judges the merit of overall instructional sequence to retain or adapt that sequence. The audience of summative evaluation is the consumer of instructional programme in contrast to the formative evaluator whose audience is the designer and the developer of the programme. A formative evaluator is a partisan of the instructional sequence and does every thing to make teaching or learning better. A summative evaluator is an uncommitted non-partisan person who is to pass judgement on an instructional endeavour. Diagnostic evaluation is in a way intensive formative evaluation concerned mainly with the placement of pupils in the instructional sequence so as to enable them to get the maximum benefit of the proper starting point for the pupils. For this, determining the entry behaviour which is a pre-requisite to learning the new unit is necessary, as is the personality pattern of students in terms of their interests, attitudes, skills, attitude etc. Diagnostic evaluation provides answer to this.

A very clear distinction is made between these three concepts by Bloom, Hastings and Madaus (7). *Summative evaluation*, according to them, is judgemental in nature. Its purpose is to appraise the end-products of the instructional efforts, whatever may be the teaching-teaching process. To distinguish it from formative evaluation, it is an end of the course activity concerned with assessment of the larger instructional objectives of a course or a substantial chunk of the course. Our public examinations, annual and term tests are all summative tests used for making judgements about students' learning or achievements. The focus of summative evaluation is on measurement of pupils' achievement and not on their improvement. Thus it is a status evaluation of students. The major function is that of grading, promoting or certification of achievement. It may take place at the end of a unit, term or a course of studies. Its emphasis is generally on measurement of cognitive behaviours

sometimes on psychomotor and occasionally on effective behaviours. Instrumentation is limited to final or summative examinations, through a weighted sample of course objectives. The average difficulty level of questions ranges from 35% to 70%. Scoring, though normally norm-referenced, can also be criterion-referenced. Reporting of scores is by objectives. Summative evaluation is thus a judgemental activity focussed on certification of students' achievement.

Formative evaluation is developmental, not judgemental in nature. Its purpose is to improve students' learning and instruction. Therefore, its major function is feedback to the teacher and students to locate errors in the teaching-learning process in order to improve it. It operates during instruction and is generally limited to assessment of cognitive behaviours. All classroom assessments which are not used for grading purposes whether these are unit tests, informal tests, questioning during teaching, home assignments, teachers' class-room observation of pupils' responses are all examples of formative evaluation. For formal testing, specially designed instruments are devised. As for judgements or scoring it is criterion-referenced not norm-referenced as in summative evaluation. Decisions made relate to steps to be taken to improve the instructional programme vis-a-vis pupils' learning. Reporting of pupils' progress is done in terms of an individual pattern of pass-fail scores on different tasks in the hierarchy of learning outcomes. Formative evaluation is, therefore, a means of determining what the pupils have mastered and what is still to be mastered, thereby indicating the basis for improvement of students' learning.

Diagnostic evaluation is in a way formative as it aims at discovering those areas of students' weaknesses that hinder his progress. This could be due to the absence of requisite entry behaviour or level of mastery in a subject, inappropriate instructional mode etc. Accordingly, the major function of diagnostic evaluation is that of placement of students.

Determining the presence or absence of pre-requisite knowledge or skills, prior level of mastery, classification of students into different modes of instruction and causes of repeated

learning difficulties all come under the purview of diagnostic evaluation. Its timing is related to placement at the commencement of a unit, term or an academic year. It has also a place during instruction when students repeatedly make mistakes and are not able to profit from instruction. It is not only limited to cognitive, affective and psychomotor behaviours as far as emphasis is concerned, but it is also concerned with diagnosis of physical, psychological and environmental factors which result in the poor performance of the students. For instrumentation, ordinary formative and summative tests like unit tests or readiness tests can be used for pre-testing. Teacher made tests, observation, checklists, standardised achievement tests, and standardised diagnostic tests are also used. A sample of weighted course objectives and a specific sample of each pre-requisite behaviour or a sample of non-educational and environmental factors are taken for sampling purposes. In testing, large number of easy items of about 65% or higher difficulty level are used. Judgements made are both norm-referenced as well as criterion-referenced. Reporting is done as individual profiles by indicating performance on subskills. Diagnostic evaluation is, therefore, focussed on identification of pre-requisite knowledge, skills and environmental factors which enable the teacher to place the pupils at the right ladder of the achievement continuum so as to enable him to achieve his maximum. If proper diagnostic evaluation is done at the appropriate time, it facilitates formative evaluation to which it is complementary as well as supplementary in providing the needed evidence. If diagnostic evaluation reflects on the learnability of a student, formative evaluation provides clues to improve his learning while summative evaluation certifies his achievement.

9. Norm-referenced, Criterion-referenced and Self-referenced Measurement

Norm-referenced measurement is the traditional class-based assignment of numerals to the attribute being measured. It means that the measurement act relates to some norm, group or a typical performance. It is an attempt to interpret the test results in terms of the performance of a certain group. This

group is a norm group because it serves as a referent or norm for making judgements. Test scores are neither interpreted in terms of an individual (self-referenced) nor in terms of a standard of performance or a pre-determined acceptable level of achievement called the criterion behaviours (criterion-referenced). The measurement is made in terms of a class or any other norm group as the function is to relate individuals' measurement to some norm group (class). The purpose is to produce response variance i.e. to see the extent to which an individual varies or differs from the performance of the group to which he belongs or belongs not.

Almost all our class-room tests, public examinations and standardised tests are norm-referenced as they are interpreted in terms of a particular class and judgements are formed with reference to the class which is considered as a type. Who is the most intelligent boy in the class? Who stood first? Who got the least marks? Is he better than 50% of the students in the class? These are questions which involve norm-referenced judgements. Such judgements use the performance of some group as a referent on the same task. We compare an individual's performance without similar information about the performance of others. That is why selection decisions always depend on norm-referenced judgements. Prediction and many placement decisions also depend on these types of judgements. A major requirement of norm-referenced judgements is that individuals being measured and individuals forming the group or norm, are alike. The conditions under which the referent (norm) was obtained and the conditions under which the original information was obtained are also assumed to be similar. Another criterion is that the referent used in norm-referenced judgements should have the minimum error so as to have reliable and accurate judgements. Unless the referent used is uptodate or recent, the comparison of an individual's performance with the group (with an outdated referent) is of no use as it would lead to faulty interpretations. Thus norm-referenced measurement pre-supposes an uptodate, reliable referent (norm group) of *like* individuals obtained from *like* conditions.

A *criterion-referenced measurement* has its origin in the

writing of objectives by Mager (9) who urged the teachers to specify a *Criterion* of acceptable performance while stating instructional learning outcome, an intended level of proficiency of the learner or a desired standard of performance. Thus in contrast to a norm-referenced measure we can reference an individuals' performance to a predetermined *criterion* which is well defined. This type of measurement is termed as criterion-referenced measurement. It is determining an individuals' status with reference to a well-defined criterion behaviour. It is an attempt to interpret test results in terms of clearly-defined learning outcomes which serve as referents (criteria). Success of criterion-referenced tests lies in the delineation of well defined *levels* of achievement which are usually specified in terms of behaviourally stated instructional objectives. According to Glaser (1963) underlying the concept of measurement of achievement lies the notion of the continuum of knowledge acquisition ranging from no-proficiency to perfect performance. It is on this continuum of knowledge that an individual's status regarding his achievement is to be determined. Unlike an norm-referenced measurement, the criterion level of the minimum acceptable performance for each objective is specified in advance in criterion-referenced tests.

Use of criterion-referenced measurement at the elementary stage where learning of basic skills and fundamental concepts is essential, is a must to a lay proper foundation for learning at the secondary stage. It does away with the unfair comparison of an individual with other children. The major difficulty is the establishment of an achievement continuum in accordance with the complexity of the skills or the concepts involved. In fact, both norm-referenced and criterion-referenced measurements have a place in the teaching-learning process and the details are discussed in a separate chapter on criterion-referenced tests.

Self-referenced measurements relate to the comparison of an individuals' performance with himself. Here the referent is neither a norm group (class) nor the criterion behaviour (expected performance) but the individual himself. It is an attempt to interpret the test results in terms of the individuals' own performance at two different times. Therefore, the referent is his

previous performance. If he scores more than his previous scores he is improving and if he scores less, then his progress is not as desired. This progress can however be determined either in terms of the rate of the progress relating to the norm group (class) or in terms of predecided or prespecified learning outcomes or performance objectives. In the former case his rank, for example, in the class or the deviation of his scores from the mean class performance indicates his progress or retardation. In the latter case his performance may be judged in terms of his comparative performance (now and then) on the instruction outcomes or level of proficiency in prespecified tasks or criterion behaviours.

The basic principle of self-referenced judgements is the acceptance of the uniqueness of the child in his learning rate. Such judgements are valuable for the weaker section of learners in preventing frustration and developing among them the sense of achievement. Ekta's position is now 10th in the class while it was 15th in the last examination. She can now solve sums on quadratic equations which she was not able to do in the last term test. She has gone down in Biology when compared to her previous performance. These are all examples of self-referenced judgements which takes into account the individuals' own capacities and efforts. All the three types of measurement—the norm-referenced, criterion-referenced and self-referenced are not mutually exclusive and should in fact be considered in the form of a triangle in which an individual is viewed differently from each angle to get the desired perspective.

10. Quantitative, Qualitative and Responsive Evaluation

The term *Quantitative evaluation* is widely used in education. In fact, the word 'measurement' is often equated with quantitative measurement. As such, quantitative evaluation refers to the assignment of numerals to an object of measurement according to certain rules. Here object means any attribute of the personality of the individual. It is historically the most potent means of evaluating students whose performance or achievement is recorded in the form of scores which are some kind of numbers. Its wider use can be attributed to the facility it offers for manipulation of numbers for comparison, selection,

grading or certification. It is applicable to all curricular areas. A major source of data for quantitative evaluation is testing, which may employ teacher-made or standardised tests for the collection of data. The scores or the marks which are normally called raw-scores are converted into various forms. These may be standard scores or any other type of comparable scores. Written examinations, unit tests, term tests, diagnostic tests, or any other form of testing, all yield data which are generally represented by numbers and is, therefore, amenable to manipulation. The validity and reliability of measuring instruments that yield numerical data can be more easily maintained or improved. Judgements which are formed on the basis of quantitative evaluation are generally norm-referenced. Quantitative evaluation provides measurement data in such a form that it is easier to interpret and to take decisions regarding teaching or learning. A major limitations of quantitative measurement is the lack of needed reliability of the measuring instrument. If marks are not considered sacred and are interpreted in the light of the quality of measuring instrument, they can be more useful. Quantitative evaluation is, therefore, the traditional mode of collecting evidences through various tests which yield numerical data that form the basis for judgement making and decision taking.

Qualitative evaluation in contrast to quantitative evaluation, is that mode of appraisal of students' achievement or growth which uses qualitative measurement or non-measurement as it is sometimes described. The object or the attribute of measurement in this type of evaluation are difficult to measure quantitatively. For example, personal and social qualities like those of intellectual honesty, civic sense, cooperativeness, patriotism, scientific attitudes etc., are some of the examples which require the use of qualitative evaluation rather than quantitative evaluation. The basic principle underlying qualitative evaluation is the assumption that the attribute may be visualized only in terms of the degree or the extent of its achievement or growth on a continuum. It may, therefore, be either in terms of a verbal description, letter grading, graphical representation, rank etc. Tools and techniques used in this type of evaluation of observation, rating scales, checklists, and such other instru-

ments which do not yield scores. Therefore, qualitative evaluation yields subjective data in contrast to the objective data that accrue from quantitative evaluation. As for the methodology of judgement making, the data are compared to the criteria or the attribute comprised by each trait. Therefore it is, in a way, criterion-referenced judgement. There is no point in comparing an individual's attainment in the affective domain with his counterparts. It is a self-referenced judgement that may be made. A major limitation of a qualitative evaluation is the lack of a well developed criterion behaviour of each trait of measurement and the subjectivity of the evaluator that looms large in such an evaluation. The qualitative evaluation is thus an attempt to interpret the evidence about pupils' growth in different fields in terms of degree, say, on a five-point scale against the criterion behaviours predetermined for the purpose.

The term *Responsive evaluation* was used for the first time by Robert Stake. According to him, responsive evaluation is oriented more to the programme activity than to the programme intents and responds to audience requirement for information. It refers to different value prospectives in reporting the success or failure of the programme. The responsive evaluator engages different observers to observe and negotiate the programme. With the help of observers he prepares narratives, graphs, portrayals, etc. Using proper negotiations with the programme personnel, he checks up the accuracy of these portrayals and reactions with the audience members, which are recorded. Suitable media are used to enhance the fidelity of communication and the final written report may or may not be prepared as agreed by him and his clientele.

Stake prefers to identify the *issues* instead of the objective or the hypotheses. According to him, issues reflect better the complexity, immediacy and valuing of the programme. The evaluator consults parents, teachers, students and other programme staff and identifies issues or potential problems. It is these issues that are basic to continue discussions, for gathering data, making observation, to conduct the interviews, test etc. All this is done to improve communication with the audience. The conventional research reporting is characterised by explicit

communication, finding relationships among various variables, analysing and interpreting in terms of variance or co-variation. In contrast, responsive evaluation is an attempt to respond in a natural way to understand what the programme is like. The people involved have practical experience of assimilating and, therefore, more efficient comprehension. Sometimes direct experience may not b e possible and is substituted by vicarious experience similar to that which members of the audience use.

Thus, data collection is not limited to one person but is collected from all persons involved. The methodology of processing the data is multi-way communication and discussion by the evaluator with the help of the observers. Reporting gives a holistic impression, portrays the complexity, the mode and the mystery of the experiences relating to the programme activities. The uncertainty and ambiguity of reports natural to such evaluation is its major limitation. The responsive evaluator may prepare portrayals of a just five-minute script or a long portrayal which involves many narratives, maps, graphs, taped conversation, photographs, exhibits etc. Responsive evaluation is, therefore, a new approach of cooperative enterprise in evaluating the activities of a programme through continuous discussion with the programme personnel and reporting of possible prospective of the success and failures of the programme.

11. Rational, Experimental and Illuminative Evaluation

Rational evaluation is the basic form of all evaluation activities. It is the most commonly used and usable form of evaluation which is based on the intelligent and enlightened opinion of the evaluator. It may also be called logical evaluation as the evaluator forms his judgement on some logic he uses as a basis. The major attributes of rational evaluation are reasoning, logic, experience, personal opinion and preferred values. The evaluator uses his own rationale, depending upon the purpose of evaluation. It has no empirical base. It is a purely subjective appraisal of a pupil, curriculum or a programme. The objectivity lies in the relevance of the criteria adopted and the rigour with which those criteria are used. Since it is not possible always to agree on such criteria, unless pre-specified or given, its objecti-

vity is ensured by involving more than one evaluator.

Rational evaluation is done by experts who are conversant with the nature of the task to be evaluated. As for its scope, this form of evaluation is applicable to almost all situations. Even in other types, initially the evaluation is done on a rational basis only. Take, for example, the relevance of instructional objectives in science; the suitability of content or topics in a history syllabus; the sequencing of units of teaching; the selection of a certain method to teach a particular concept; to organise learning experiences around pupils' interest, topics or activities; giving weightage to different forms of questions; estimating the difficulty level of a question; discouraging overall options etc. are all matters of rational evaluation to start with, although each of them can be empirically verified.

All the above examples indicate that rational evaluation is basic to all components of the teaching-learning process. The basic tenet of rational evaluation is the finding out of the logical relevance of the material on the basis of experienced expertise. The source of data is the evaluator's own fund of knowledge and skills and the logic of the discipline. The methodology of data processing rests on the criteria evolved in advance. Judgements involve reflection on the predetermined nodal points. To what extent ends and means are articulated is the focus of rational evaluation. Its chief advantage is the economy of time and expert opinion. Its limitation is its subjectivity in forming judgements. As compared to other modes of evaluation this is by far the most popularly acclaimed method of evaluation used in most situations relating to the educational process. However, its wider applicability cannot be considered as an indicator of its greater validity or reliability.

Experimental evaluation and *illuminative evaluation* have emerged from two distinct paradigms. Each has its own strategy, focus and assumptions. The first one which is more dominant is based on 'classical' or 'Agricultural-Botany' (3) paradigm derived from the experimental and mental testing tradition in psychology and it therefore utilises the hypothetico-deductive methodology. According to Parlett and Hamilton this type of evaluation (4) "students—rather like plant crops—are given pretests and then

submitted to different experiences. After a period of time, their attainment is measured to indicate the relative efficiency of the methods used." This type of approach yields objective numerical data that permit statistical analysis. Variables like I.Q. test scores, attitude rating, personality profiles etc. are codified and processed to indicate the efficiency of the new curricula, media or methods.

Intelligence testing, achievement testing, development of the attitude scales, interest inventories and personality inventories are all examples of experimental evaluation. It involves controls of variables in a predesigned experiment to meet the dictates of the paradigm. Testing, which takes place generally before and after the research design, is the prominent form of this type of evaluation. The evaluator himself is the source of data which are gathered by him mainly through written tests. A well set methodology is used to process the data. Judgements are generally norm-referenced.

A major limitation of this type af evaluation is the difficulty of controlling variables, expensiveness, time-consuming data-collecting exercise, absence of provision for adapting the built-in premise according to changed circumstances, over reliance on quantitative information, insensitivity of statistical generalisations to local and unusual effects, failure to articulate the concern of participants, developers, sponsors and other interested personnel. Experimentation evaluation is thus not vulnerable to manifold extraneous influences which narrow down its claimed empirical reality and tends to make it more centralised and bureaucratic in its decision making process.

Illuminative evaluation is based on a contrasting paradigm which differs from the classical paradigm in research style and the methodology of processing. It is called the Social-Anthropological, psychiatory and participant observation research in sociology (5). The term illuminative research is drawn from Trow (6) (1970). This term takes care of the wider context in which a programme functions. Description and interpretation rather than measurement and prediction is its concern. Its aim is to study how a programme operates and is influenced by the various school-situations it attempts to document like a parti-

cipant in the scheme, besides discerning and discussing the significant features and recurring concomitants and critical processes. "Thus it addresses and illuminates a complex array of questions which have bearing on the process of education and enlighten the innovator and other interested parties in identifying those elements or procedures which seem to have had desirable results (7)."

The instructional system and the learning milleu are two concepts central to the understanding of illuminative evaluation. Evaluation concentrates on the 'process' of learning rather than on the 'outcomes' of learning. Those who are involved in the process and the conditions and constraints under which a project works in the class-room, are the parameters of this type of evaluation. Therefore, the perception account of the learning milleaux is essentially like travellers' tales (8). Its place in the teaching-learning process is obvious especially in the non-scholastic area of pupil growth which demands evidence from all types of participants, students, teachers, parents, peers etc. The tools used are mainly observation, discussion, interview schedule, opinionnaire etc. The methodology of processing the data is subjective estimates and face-to-face discussion with the participants. Judgements are descriptive and consensus-based. A discussion would lack certainty because of the difficulty of the analysis of complex situations in the process of education. Illuminative evaluation thus represents a sociological approach based on an appraisal of the complex interaction of various parameters having a bearing on the pupils' learning and forming judgements on the basis of the consensus of all those who are participants in the teaching-learning process.

12. Objective-based, Goal-free and Pay-off Evaluation

This tripartite classification is based on the criteria of the success of a programme. Evaluation theorists like Popham (12) classify various models of evaluation into three categories; the goal attainment models; the judgemental models emphasising extrinsic or intrinsic criteria and decision facilitation models. The *Objective-based evaluation* comes under the first category in which emphasis is on the determination of the degree to which an instructional objective is achieved. It is more ancient

in lineage and is associated with the name of Ralph W. Tyler (14) as reflected in the well-known Eight-Year Study of the 1930s. This approach envisages careful formulation of goals which are transformed into measurable or behavioural objectives. Objective-based evaluation reflects the adequacy or inadequacy of the instructional programme, depending upon the degree to which the predetermined goals are achieved. This type of evaluation is widely used in class-room instruction and evaluation as also in many curriculum projects of today.

The basis of data collection is the intended instructional objectives and, therefore, tools and techniques are also geared to collect data in terms of the pre-established objective or goals. Analysis is done in terms of objectives as also their interpretation. The methodology of forming judgements and taking decisions rests on the level of attainment of the goals specified in advance. Its major limitation is the neglect of collateral efforts which may hinder the attainment of the objectives besides the validity of the objectives themselves, which is of considerable import. Nevertheless, the thrust of objective-based evaluation remains on the degree of attainment of pre-specified instruction goals.

Goal-free evaluation is the term coined by Michael Scriven (15) as a sequel to the inadequate explanation offered by goal attainment models of evaluation, for reflecting on the programme whether intended or unanticipated. It takes care of the results accomplished and is not concerned with the rhetoric of instructional objectives. It tries to ignore whatever knowledge of project goals that is likely to contaminate the programme effects. It encourages the evaluator to take care of a wider range of outcomes of a programme than the objective-based evaluator who develops a 'tunnel-vision' due to focussing all the time on judging the attainability of the project goals.

Scriven does not consider goal-free evaluation as a replacement of goal-based evaluation but a supplement to the objective-oriented framework. According to him, goal-free evaluation may function internally or externally. The goal-free internal evaluator, who is a member of the curriculum developer's team, can assess the programme outcomes while the goal-free external

evaluator who is uncommitted to the programme can also do summative evaluation. The Humanities Curriculum Project of U.K. directed by Stenhouse (16) is an example of goal-free evaluation: Although many of the Indian projects are goal-free in their origin, they are no longer concerned with the various programme outcomes, consequences or effects.

The place of goal-free evaluation cannot be over-emphasised. Most informal class-room evaluations are goal-free and are concerned more with instructional effects than with instructional objectives. The source of data is not only the test scores in this case but also other tangible impact and collateral effects. Therefore, the opinion of students and other teachers matter a lot besides the formal evaluation. The processing of data and methodology of forming judgements is in accordance with the conditions, constraints and other factors which have a bearing on programme effects. The decisions relate not to the attainment of objectives but to the observed outcomes of the programme. A major limitation of goal-free evaluation is the lack of accepted criteria or goals against which the evidence collected could be validated. Thus goal-free evaluation should be considered as complementary to, rather than a substitution for, objective-based evaluation.

Pay-off Evaluation is the outcome of the judgemental model of evaluation, which involves extrinsic or intrinsic criteria. In both, professional judgements are involved. The evaluator influences so much that it may turn out to be a favourable or unfavourable evaluation. Such judgements can use intrinsic criteria or extrinsic criteria. The former are generally referred to as process criteria and the latter as product criteria (not product evaluation). While evaluating a class, internal criteria may be the proper coverage of objectives and content or the weightage given to the form of questions. On the other hand, the extrinsic criteria may be the efficiency of the test as judged by students' performance on the test when administered. This term 'pay-off evaluation' coined again by Scriven, is based on the judgemental model of evaluation involving extrinsic criteria. These criteria relate to the programme effects in contrast to the intrinsic criteria which are concerned with the internal characteristics of an instructional programme. According to him,

intrinsic factors can also not be ignored and thus he is not disdainful of instrinsic evaluation. In fact, he suggests hybrid evaluation which combines both intrinsic and extrinsic evaluations.

In most of the empirical investigations, extrinsic criteria form the basis of judgement and are therefore pay-off evaluation. Data are gathered by personnel uncommitted to the programme. Decisions have to be taken in the light of criteria external to the programme. Compared to the other two types, pay-off evaluation may or may not be goal-oriented but this type of evaluation leans heavily on the assessment of dividends in terms of extrinsic criteria.

13. Product, Process and Product-cum-Process Evaluation

The dichotomy of product and process seems to have originated in the science laboratory or a wood workshop where performance of a task is involved. *Product evaluation* relates to finding out the worth of a production, result, outcome or a yield. To distinguish it from other types of produce, this term is restricted to the psychomotor effects or consequences. In other words, product evaluation refers to the appraisal of skill performance which yields some results. Product evaluation may be restricted to the completed product or it may be extended to the assessment of techniques or the process of production. The product of performance and process of performance are two attributes of an end-product resulting from a psychomotor activity. Accordingly, the terms product evaluation and process evaluation are in use. However when both product and process are to be assessed, we use the term product-cum-process evaluation.

Such type of evaluation is applicable to practical examinations in science subjects and in assessment of performing arts like painting, dancing, sculpture, typing, wood-work, or socially useful, productive work. The product of performance may be identification of a slide or a specimen or its correct classification, a piece of improvised apparatus, a summary or findings of an experiment, a completed painting or a piece of sculpture or an article of wood-work. For assessment purpose the three basic criteria are: the quality, the quantity and the speed. Accordingly, quality scales, rating scales or timing devices are

used to collect the needed evidence about the product being appraised. When applied to the cognitive domain, product objectives are listed in terms of major concepts and generalisations arrived at through the inquiry processes.

Process evaluation refers to the various steps, procedures or inquiry processes undertaken to develop the product intended. It means the process of performance that leads to a particular product of performance. In the production of a model (Product) one might have to use various processes like the selection of appropriate tools and instruments, their efficient use, observation say under the microscope, relating the observed with the intended, are the various procedural or operational steps representing the process of performance. The product is the model. Likewise A.A.A.S Plan (11) highlights the role or the processes of scientific inquiry in the understanding of science at the elementary stage. In the lower primary grades, the processes of sciences recommended are observing, using the space-time relationship, using numbers, measuring, classifying, communicating, predicting and inferring. In the upper primary grades, the processes recommended in addition, are formulating hypotheses, controlling variables, interpreting data, defining operationally and experimenting. These are termed as process objectives and evaluation in terms of process objectives is called process evaluation.

The dichotomy of process and product is more apparent than real. In fact, conceptualisation of both as complementary to each other is necessary. Thus *product-cum-process evaluation* is the term coined for combined evaluation. This does not, however, mean that both have equal significance. What is important is the evaluation of both the process of performance as well as the product of performance. In certain cases, the process is more important than the product of performance. For example while analysing a salt to identify it, the procedural steps in sequence i.e. the process of performance is more important than mere identification of the salt. On the other hand a piece of model or a painting can be evaluated without any regard to the process used in making of these items because the quality depends on the efficiency of the process used. Likewise

there are situations where both product and process are significant from the point of view of evaluation. For example in preparing a slide for examination, we require specific sequential steps which must be appraised as also the final product i.e. the quality of the slide prepared. The time taken to yield an end-product and the quality of the product are essential for product evaluation. Process evaluation demands on-the-spot observation of the process of performance and in some cases even the written record supplements the evidence about efficiency of the process.

For data collection use of a checklist can be made to record, on the basis of observation, the presence or absence of certain component skills or integrated skills. The methodology of judgement making is, by and large, criterion-referenced. Quality scales, rating scales or checklists used, presuppose the requisite quality or standard of performance against which the product-process can be appraised on the basis of comparison. Analysis of the processes of performance helps the teacher to locate areas or skills that cause difficulty in learning. It is on the basis of the process evaluation that the evaluator can pass a better judgement on the product of performance. Process evaluation has a great diagnostic value on the basis of which instruction can be improved, as also the standard of performance of students. Process evaluation is formative in nature while product evaluation is summative in its character.

14. Micro, Macro and Elemental Evaluation

Micro evaluation represents that process of collecting evidences and forming judgements that relate to a very small sample or content or area of operation. A major attribute of a micro-evaluation programme is its highly retricted sample which forms the basis for evaluation. Like micro teaching, micro-evaluation is a scaled-down evaluation encounter in which the complexity of the evaluation process is simplified in size and time. This can be done in two different ways. One, that all steps of evaluation may be taken to evaluate a product or process but the sample size is reduced to the minimum. For example, a teacher may take a small unit, a module or even a concept and then

evaluate the student on that portion only. Development of a concept can be tested at various levels of instructional objectives ranging from simple recall to the ability to hypothesise and evaluate. The other approach to evaluate students could be to take a particular objective and evaluate them on various behavioural outcomes like ability to recall, recognise, translate, interpret, extrapolate etc. In its extreme form, only one specific intended learning outcome, say the ability to interpret, may be taken and a pupil's achievement guaged with respect to that specific objective only.

Micro evaluation is, therefore, a more intensive probe with respect to a limited sample of content or an objective. When the time is short, the evaluator can judge the total sample on the basis of limited and similar sample. But it is, in fact, the teacher who needs more often micro-evaluation when he observes repeated failure of students to understand a concept or achieve a particular objective. Such an evaluation is focussed more on feedback of results rather than on grading of achievement. It is, therefore, formative in its approach. Data collection is based on a comparatively more controlled variable and is, therefore, more reliable. Judgements are generally criterion-referenced and decisions are directed to improve learning and teaching.

Macro evaluation, sometimes called *mega evaluation*, is a large-scale evaluation effort which has extensive coverage of content, objective or sample size. It is an attempt to interpret data in terms of a large sample. Extensiveness may vary vertically ranging from one class to all classes in a school, or it may be horizontal assessment of the curricular and non-scholastic traits of students from a school or a state. Macro-evaluation, unlike micro-evaluation, provides a megascopic view of students' growth. Public examinations, standardised tests, internal assessment programmes, a crash programme for evaluating textbooks in various subjects, are examples of different kinds of macro-evaluation. The focus of such evaluation is on measurement of growth and the judgements formed are generally summative in nature. Decision focus is norm-referenced comparison or status position.

In fact, extensiveness or intensiveness of evaluation effort in micro and macro evaluation can be viewed on a continuum ranging from measurement of a single concept and ability in class (micro) to that of a nationwide programme of evaluation like that of the Talent Search Examination of the N.C.E.R.T. or the U.P.S.C. Central Services Examinations. The purpose varies from diagnosis and improvement of learning to that of grading and certification of students, curriculum or a project as we move from one end of micro-evaluation to the other end of macro-evaluation.

Elemental evaluation is an attempt to form judgements about a single but specific and predetermined element or attribute of a person, text, curriculum or a programme. For example, a history textbook may be evaluated only from the point of view of national integration or a science text from the point of view of investigatory approach. Likewise, a curriculum can be evaluated from the point of view of the requirement of the learner or social needs etc. A programme may be evaluated in terms of its context, input, process, product impact or collateral effects. Similarly, pupils' evaluation may be done only to judge their civic sense, cooperativeness, expression, handwriting or spellings. The focus is only on one element or component of learning, curriculum, text, a programme or a project.

Elemental evaluation differs from micro-evaluation in that it may give a microscopic or megascopic report depending upon the sample used. The one refers to the scope aspect of evaluation which may be elemental or comprehensive, whereas the other refers to the nature and scope of the evaluative effort to be made with respect to that aspect. Basic to elemental evaluation is the identification of the expected behavioural outcome i.e. the criterion behaviours or the indicators of success. This type of evaluation can be undertaken, depending upon the preferred objectives, values, aspects or elements. Elemental evaluation is thus an attempt to interpret the evidence about a person, programme or a project in terms of a single aspect about which judgements are formed and decisions are taken.

15. Programme Evaluation, Curriculum Evaluation and Pupil Evaluation

Programme evaluation refers to all the activities or steps taken to judge the effectiveness of a programme or a project. A programme may be a project only or it may consist of more than one project in the total programme. A programme of non-formal education may for example be composed of different projects like functional literacy, adult education, development of literature for the neo-literate, evaluation of pupils' growth etc. A programme is, therefore, to be evaluated in terms of its objectives as well as the means adopted for the achievement of those objectives. Since every programme is conceived in its own context and is born in its own soil, it has to be evaluated in the same context. Likewise, inputs provided, procedures adopted for implementation, materials prepared, outcomes or products accrued and impact made, are the different aspects which come under the purview of programme evaluation. Thus terms like context evaluation, input evaluation, process evaluation, outcome evaluation, product evaluation, impact evaluation, materials evaluation, curriculum evaluation, pupil evaluation etc. are all used under the umbrella of programme evaluation. A programme evaluation, therefore, may include curriculum evaluation as well as pupil evaluation because one of the criteria of effectiveness of a programme may be the effectiveness of the curriculum or the pupils' growth as a result of programme impact.

Data collection involves all the four major techniques, namely observation, opinion, analysis and testing. Methodology of judgement-making varies with the programme intents or with reference to a comparative programme. Decisions generally lead to programme acceptance, programme modification or programme termination.

Curriculum evaluation is a part of programme evaluation or an independent project when the programme itself is of developing a curriculum. Since curriculum is a comprehensive term used to include curriculum objective, curriculum content, curriculum methodology and evaluation, it involves evaluation of the objectives syllabus, textual material or other materials, methodology of

CHAPTER III

BASIC STATISTICAL CONCEPTS IN MEASUREMENT

Statistics is a basic tool of measurement, evaluation and research concerned with gathering, analysing and interpreting numerical data by using various processes or mathematical techniques. This word is also used to describe the collected numerical data. On the basis of individual observations, group characteristics are abstracted to form generalisations by means of statistical data. The average family income, the average 6th grade boy is 11 years old, the average marks secured in Mathematics by 8th class students are all statistical concepts. For a research worker, there is need for identifying the facts relevant to the questions or the hypothesis to be tested, the mode of collecting, organising and analysing observations, assumptions underlying the methodology of processing data and the validity of conclusions. Thus systematic observations and description of characteristics of objects and events, finding relationships between various variables, making generalisations and predicting future occurrences are basic to research. Measurement is the universally accepted and most precise method of describing and assigning quantitative values to the characteristics of objects and events.

It is expected that after going through this chapter the reader will be able to

(a) differentiate between different types of scales, parametric and non-parametric data, descriptive and inferential statistics etc.
(b) organised data into frequency distribution, cumulative

frequencies etc. for making scores meaningful;

(c) calculate measures of central tendencies, standard deviation, variance, relationships like co-efficient of correlation and use them to interpret measurement data;

(d) understand the concept of normal probability curve and its application;

(e) compute reliability co-efficient using different methods, the standard error of measurement and standard error of estimate;

(f) appreciate the role of basic statistical concepts in understanding measurement data and their meaningful interpretation for comprehending the process of measurement and evaluation.

1. Measurement Levels

The nature of variables and the precision of measuring instruments determine the level of precision and sophistication of the measure or the scales. Following are the four scales in use:

1.1. *Nominal Scale*

For example in an office we have:

	Female	Male	Total
Lower division clerks	29	9	38
Upper division clerks	15	5	20
Assistants	10	3	13
Superintendents	04	2	06
Section Officers	02	1	03
Total	60	20	80

—Objects (Persons) are categorised.

—Each individual can be a member of one set only.

—All members of the same set have same defined characristics.

—They are non-orderable i.e. cannot be ranked.

—Counting is the only feasible method of quantification for statistical analysis.

Such a scale which indicates or describes difference between things by assigning them into nationalities, genders, educational levels, occupations etc. is called a *nominal scale.*

1.2. *Ordinal Scale*

For example the height of a group of girls is as under:

Girls	Height in cms.	Difference in cms.	Rank
Rita	160	—	1st
Nita	158	2	2nd
Sita	155	3	3rd
Gita	150	5	4th
Mita	142	8	5th

—Ordering of things into more than or less than adjacent things.
—Ordering is expressed in terms of rank in the group.
—No absolute values and the difference in adjacent ranks may not be equal.
—Ranking spaces them equally but they may not be equally spaced.

Such a scale which not only indicates that things differ but also that they differ in amount or degree is called an *ordinal scale.*

1.3. *Interval Scale*

It is based on equal units of measurements. It indicates the magnitude of a given characteristic or property. A difference between a score of 75 and 76 is assumed to be equivalent to the difference between 50 and 51 and 16 and 17. Since it cannot measure the complete absence of a trait, its main limitation is the lack of true zero. Psychological tests and inventories are *interval scales.*

1.4. *Ratio Scales*

Like the interval scale it has equal interval property. In addi-

tion it has a true zero. For example the zero point on a centimeter scale indicates complete absence of length or height. Numerals of ratio scale can be added, subtracted, multiplied, divided and expressed into relationships e.g. 2 gms. is one half of 4 gms. or twice the 1 g.

2. Data

2.1. *Parametric Data*

It is measured data. Parametric statistical tests assume that data are normally or nearly normally distributed and are applicable to both interval and ratio scaled data.

2.2. *Non-parametric Data*

This type of data are either counted or ranked. They are also called distribution-free tests and do not rest on the assumption of normally distributed populations. To summarise:

(a) Level	1	2	3	4
(b) Scale	Nominal	Ordinal	Interval	Ratio
(c) Process	Counted and classified	Ranked in order	Measured equal internal, No true zero	Measured equal intervals True Zero Ratio relationship
(d) Data Treatment	Non-Parametric	Non-Parametric	Parametric	Parametric
(e) Appropriate test	Chi-Square Median Sign	Spearman's rho (ρ) Mann-Whitney Wilcoxon	t-test, analysis of variance analysis of covariance, Factor analysis, Pearson's r	

3. Descriptive and Inferential Analysis

When statistical analysis is limited to generalisations applicable to a particular group of individuals observed, it is descriptive analysis. Data describe one group and that group only and no conclusions are extended beyond this group. Simple action research is an example of this type.

When analysis involves a sampling process to select a small group which is assumed to be related to a larger group from which the small group is drawn, it is inferential analysis. The small group is called the sample while the large group is called population. Thus drawing conclusions about population on the basis of observations of the samples is the purpose of inferential analysis.

4. Statistic

When statistic is used as a singular it is a measure based on observations of the characteristic of a sample (mean, standard deviation etc.). A statistic computed from a sample may be used to estimate a *parameter*, the term used to connote a value corresponding to the population from which the sample is drawn.

A sample must, therefore, approximate to the larger group i.e. the population. Only then it is possible to estimate the characteristics of the population by analysing the characteristics of the sample. If sampling is not done carefully, the research worker should restrict the findings to the group observed and should not apply them to other individuals or groups. The statistical theory of sampling is complex and the details are not warranted in this introductory treatise.

5. Organising Data for Making Scores Meaningful

Let us take the following scores of 10 students written either roll-number-wise or alphabetically.

1. Amita	= 31	6. Lalita	= 36
2. Anita	= 16	7. Manjeeta	= 29
3. Babita	= 35	8. Ranjeeta	= 35
4. Kavita	= 28	9. Sarita	= 26
5. Gita	= 38	10. Vanita	= 26

From this type of data it is not convenient to know the highest score, lowest score, average score and how the score of Kavita (28) will compare with others in her class.

5.1. *The Array*

The same scores when arranged in say, descending order, as 38, 36, 35, 35, 31, 29, 28, 26, 26, 16, become more meaningful and at a glance we can know the highest score (38), the lowest score (16), the average score (30) and the performance of Kavita with 28 marks (just below average).

5.2. *The Rank*

These scores can now be given ranks as 1, 2, 3,.....10 and the teacher can assign ranks without giving the raw scores. When two scores are the same then the ranks in this case can be calculated by dividing the summed up ranks of the two students and giving the same rank to both the students as under:

Score:	38	36	35	35	31	29	28	26	26	16.
Rank:	1	2	3.5	3.5	5	6	7	8.5	8.5	10.

5.3. *Frequency Distribution* (F.D.)

When the scores are more than 20 it becomes difficult to rank and is time consuming. In such cases score are grouped into intervals of 2, 3, 5 or 10 depending upon the range and number of scores. Scores are then tallied in each interval. Tallies are counted and the number is put under the frequency column. The sum of the tallies/frequencies should equal the number of scores. The following example illustrates the frequency distribution of 20 students' scores on a unit test in Biology.

Scores: *38, 21, 35, 39, 31, 33, 32, 26, 29, 28, 24, 34, 25, 32, 30, 33, 31, 30, 28, 23*

TABLE 1: *Frequency distribution of scores of 20 students in Biology*

Interval	Tallies	Frequency
39—40	/	1
37—38	/	1
35—36	/	1
33—34	///	3
31—32	////	4
29—30	///	3
27—28	//	2
25—26	//	2
23—24	//	2
21—22	/	1
	N=20	

Steps

(a) Number of intervals can be set between 20 and 40 by dividing the range (39−21=18) by 10 to estimate the size of interval 18÷10=1.8. Therefore, in this case an interval of 2 was used.
(b) Establish the highest interval (39-40) and other intervals down through the lowest interval containing lowest scores (21-22).
(c) Tally the individual scores in each interval. Count the tallies and write the number in the frequency column.
(d) Add the frequency column to check that the sum equals the number of scores in the collection.

A score interval of odd number of units may be preferred to even units because mid-point is a whole number and not a fraction and is, therefore, convenient for computation since all scores are assumed to fall at the mid-point.

Example: Interval of 6 (even) 7, 8, 9, 10, 11, 12 (mid-point 9.5)
Interval of 5 (odd) 7, 8, 9, 10, 11 (mid-point 9.0).

6. Statistical Measures

To describe and analyse the data in a meaningful way you should be able:

1. to calculate the measures of
   (a) central tendencies or averages like mean, median and mode,
   (b) dispersion or spread like deviations, variance and standard deviation,
   (c) relationship like that of coefficient of correlation.
2. to use these measures to the given data for meaningful interpretation of test results.

6.1. *Measures of Averages* (Central Tendencies)

(i) *Mean* (M)

Arithmetic average, grade point average and Arithmetic mean,

all represent the mean value and are very useful in classroom measurement. A widely used index of average is the Arithmetic mean which is calculated by adding the separate scores and dividing the sum by the number of individual scores in the collection.

Ungrouped data

$$\text{Mean} = \frac{\text{Sum of scores}}{\text{Number of scores}} \text{ OR } \frac{\Sigma X}{N} \quad (1)$$

where ΣX = Sum of scores in a distribution
In Table 1 sum of 20 scores is 602.
Thus N = Number of scores, and $\Sigma X = 602$.

$$\text{Mean} = \frac{602}{20} = 30.1$$

TABLE 2: *Calculation of Mean from Grouped data*

Interval	X (Mid-point)	f (Frequency)	fX
39—40	39.5	1	39.5
37—38	37.5	1	37.5
35—36	35.5	1	35.5
33—34	33.5	3	100.5
31—32	31.5	4	126.0
29—30	29.5	3	88.5
27—28	27.5	2	55.0
25—26	25.5	2	51.0
23—24	23.5	2	47.0
21—22	21.5	1	21.5
		N=20	$\Sigma fX = 602.0$

$$\text{Mean} = M = \frac{\Sigma fX}{N} = \frac{602}{20} = 30.1 \quad (2)$$

Here, by chance, the mean is the same for grouped and ungrouped data. Generally, there is small difference between the means calculated from both grouped and ungrouped data methods. The concept of the mean is useful not only in understanding the performance of a group but also because it is used as a basic statistic to compute other statistics like standard deviation etc.

(ii) *Median* (Md)

Median is a score point (not necessarily a score) in an array, below and above which one half of the scores fall. It is a measure of position rather than of magnitude and can be found without calculating, by inspection if the scores are arranged in ascending or descending order. A median can be calculated as under:

(a) If the number of scores is odd:

Use formula:

$$\text{Median} = \text{Md} = \text{Mid-score} = \left(\frac{N+1}{2}\right)\text{th score is an array.} \quad (3)$$

Example

Scores

5 — 2 Scores above

4 ↑

3 — Median OR $\frac{5+1}{2}=3$, the 3rd score from below i.e. 3.

2 ↓

1 — 2 Scores below

(b) If the number of scores is even:

Use formula:

$$\text{Median} = \text{Md} = \text{Mid-point between the two middle scores.}$$

$$= \text{Mid-point between } \frac{N}{2}\text{th and } \left(\frac{N}{2}+1\right)\text{th score}$$

Scores

6 — 3 Scores above

5 ↑

4 |

Median = 3.5 OR $\frac{6}{2}=3$ Middle scores will be 3 and 4.

3 |

2 ↓ OR $\frac{3+4}{2}=3.5$

1 — 3 Scores below

The median is useful and is used widely in classroom work because

(a) it is easily found by just ordering the scores,
(b) it divides total scores into two halves which depict the average and examinees can understand it better,
(c) it is not influenced by extreme scores at either end of the distribution and becomes a more realistic measure.

Example

Scores	Scores
5	24
4	5
3—Median	Median—3
2	2
1	1

In this case Mean is 3 in one and 7 in the second whereas median is 3 in both the cases.

(iii) *Mode* (MO)

Mode is that score which occurs most frequently in a distribution. It is located by inspection rather than by computation. In grouped data, the mode is assumed to be the mid-score of the interval in which the greatest frequency occurs.

Ungrouped data	*Grouped data*	
Scores	Interval	Frequency
8	19—20	8
7	17—18	7
5 5 } mode 5	15—16	15—Mid-score of (15—16)=15.5 (Mode)
4	13—14	4
3	11—12	3

Examples

—Modal age of 5th class students is 11 years
—Modal size of gents' shirts is 38 cm.

This means that there are more 11 year-old students in class V than of any other age. Likewise, there are more people with 38 cms' collar size than any other size.

Distribution may be bimodal or multimodal. For example,

the number of students travelling by D.T.C. on working days may be more between 8 and 9 A.M. and then between 5 and 6 P.M. thereby giving a bimodal distribution.

6.2. *Measures of Dispersion or Spread*

A measure of dispersion or spread is an index that describes the extent to which scores are spread out from the average score. In other words, it indicates the heterogeniety or homogeniety of the scores and provides a basis for comparison of different collections of scores.

(i) *Range*

It is the *simplest measure* of dispersion and is simply the difference between the *highest* and the *lowest score*. Although it is useful for making simple interpretation, it is unreliable and misleading as it is based on two most atypical scores.

Examine the following data showing scores of two groups on the same test.

Group I	Group II
20	20
5	16
4	11
3	7
2	3
1	1
Range=19	Range=19

Although the range is the same in both, the scores in group I are quite close while they spread across the entire range in group II.

(ii) *Deviation from the Mean* (Small x)

A deviation score is the score expressed as its distance from the mean.

$x=(X-M)$ where X=Score; M=Mean.

Deviation Score is positive (+) if the Score (X) falls above the mean (M).

Deviation Score is negative (−) if the Scores (X) fall below the mean (M).

Example

Pupil	Score	Grade	Pupil	Score	Grade
Amita	10	A	Lalita	9	B
Anita	9	B	Manjeeta	8	C
Babita	8	C	Ranjeeta	8	C
Kavita	7	D	Sarita	8	C
Gita	6	F	Vanita	7	D
	$\Sigma X=40$			$\Sigma X=40$	
	$N=5$			$N=5$	
	$M=\frac{40}{5}=8$			$M=\frac{40}{5}=8$	
	$Md=8$			$Md=8$	

Mean and median are the same, yet Group I is compared to heterogeneous Group II which is more homogeneous.

Using the above scores to compare the two groups let us find out the deviation scores.

Group I		Group II	
Score x i.e. (X−M)		Score x i.e. (X−M)	
10	+2	9	+1
9	+1	8	+0
8	0	8	−0
7	−1	8	−0
6	−2	7	−1
$\Sigma X=40$	$X=0$	$\Sigma X=40$	$X=0$
$N=5$		$N=5$	
$M=8$		$M=8$	

Thus sum of scores' deviations from the mean equals zero:

$$(X-M)=0 \qquad x=0$$

Therefore, mean is that value in a distribution about which the sum of the deviation scores equals zero.

(iii) *Variance* (σ^2)

Variance is the sum of the squared deviations from the mean divided by N. If we square each deviation from the means, i.e.

$-1, -2, +1, +2...$

We get a positive score. These scores can be summed up and divided by N and the mean of the squared deviations computed.

$$\text{The Variance} = \frac{\Sigma(X-M_x)^2}{N} \text{ OR } \frac{\Sigma x^2}{N} \text{ where } x=(X-M_x)$$

and M_x=Mean of X scores (4)

Thus variance is indicative of how all scores in a distribution are distributed or dispersed about the mean. This concept is employed in a number of statistical interpretation of data.

(iv) *Quartile deviation*

Since extreme scores affect the size of the range it is difficult to interpret unless these extreme scores are neglected. This is done in quartile deviation. If range is computed over the middle only instead of the whole range we can avoid the effect of extreme scores on our results. Quartiles like the median, are points on the continuum and may not always fall at points occupied by scores.

Example

		Middle 50% of scores								
Distribution—A:	5, 6, 8	9	9	9	11	12	14	15	17	20
	First Quartile							Third Quartile		
Distribution—B:	5, 5, 5	6	6	6	7	7	9	10	10	10

(a) There are 12 scores in each distribution.

(b) Isolate 50% of scores (i.e. 6) in the middle.

(c) First quartile or Q_1 is the point which separates the first quarter from the middle i.e. the point below which lie 25% of the scores.

(d) Third quartile or Q_3 is the point which separates the third quarter from the middle i.e. the point below which lie 75% of the scores.

(e) Third quartile (Q_3) in Distribution—A falls between 14 and 15 i.e. *14.5*.

(f) First quartile (Q_1) in Distribution—A falls between 8 and 9 i.e. *8.5*.

(g) Interquartile range $= Q_3 - Q_1$ OR $14.5 - 8.5 = 6.0$.

(h) Semi-quartile range $= \frac{6}{2} = 3$. It is also called Quartile Deviation.
Therefore,
Quartile deviation $Q = \frac{Q_3 - Q_1}{2} = \frac{14.5 - 8.5}{2} = \frac{6}{2} = 3.0$

(i) For Distribution—B

Quartile deviation $Q = \frac{9.5 - 5.5}{2} = \frac{4}{2} = 2.0$.

Thus the difference in the quartile deviation of Distribution—A i.e. 3.0 and Distribution—B i.e. 2.0 gives a clearer picture now of the difference in the spread of scores of the two distributions.

(v) *Standard deviation* (σ) (ungrouped data)
Standard deviation is based on the difference between each score and the mean. The larger the differences are, the more spread out the scores are and the larger will the standard deviation be. It is the positive square root of the variance as shown in the formula below:

$$\text{S.D.} = \sqrt{\frac{\text{Sum of (each score} - \text{the mean) and squared}}{\text{Number of scores}}}$$

$$\text{OR} = \sqrt{\frac{\Sigma x^2}{N}} \qquad (5)$$

Example

Score	x	x^2
42	+ 2	+ 4
41	+ 1	+ 1
40	0	
39	− 1	+ 1
38	− 2	+ 4
		$\Sigma x^2 = 10$

Variance $= \frac{10}{5} = 2$ Standard deviation $= \sqrt{\frac{10}{5}} = 1.414$.

Simple formula for computing S.D.
The easiest formula which approximates to the above formula and can be used by every classroom teacher is given here:

$$\text{S.D.} = \frac{\text{Sum of the highest} - \text{Sum of the lowest}}{\frac{1/6 \text{ of scores}}{\frac{1}{2}(N-1)} \quad \frac{1/6 \text{ of scores}}{\frac{1}{2}(N-1)}} \qquad (6)$$

where

N is the number of scores in a group.

Example

Scores	x	x^2
13	+ 6	36
12	+ 5	25
11	+ 4	16
10	+ 3	9
9	+ 2	4
8	+ 1	1
7	− 0	0
6	− 1	1
5	− 2	4
4	− 3	9
3	− 4	16
2	− 5	25
1	− 6	36
	x^2	= 182

Using Formula

$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$; we get $\sigma = \sqrt{\frac{\overline{18}}{13}} = \sqrt{14} = 3.74$

Use of a simpler formula for calculating standard deviation:

$$\text{S.D.} = \frac{1/6\ H - 1/6\ L}{\frac{1}{2}(N-1)}$$

Total scores are 13. The 1/6th scores come upto two top and two bottom scores i.e. 13+12 and 1+2

$$\text{S.D.} = \frac{(13+12)-(1+2)}{\frac{1}{2}(N-1)} = \frac{25-3}{\frac{1}{2}(13-1)} = \frac{22}{6} = 3.67$$

For most of the classroom tests this estimate is sufficiently accurate and quite easy to calculate SD by this formula.

There are some important uses of standard deviation.

It can be used:

(a) to describe the relative homogeniety of the two classes to whom same or comparable tests are given,

(b) to determine the composite score based on addition of two or more scores, e.g. on two or more unit tests. The unit test with larger S.D. determines a student's standing in a class on the composite score obtained, and

(c) to derive comparable scores for a given class on two or more tests thereby helping the teachers to describe the relative quality of performance of students on different tests in relation to the mean of the distribution.

Example

Test I	Test II
Mean=30	Mean=40
S.D. = 5	S.D. = 8

A student's score of 35 on Test I will be at the same relative level on Test II if he gets 40+8=48 because 35 is one S.D. above the mean (30) as 48 is 1 S.D. above the mean (40) of second test.

The scores expressed as S.D. distances from a mean are called Z scores.

7. Derived Scores

Raw scores cannot be used to compare an individual's performance with a group on the same test or on a different test because the unit on one test may not indicate the same increment of performance as the same unit on another test. For this reason raw scores are converted into derived scores which can be interpreted more easily and are more comparable from one test to another and from one group to another.

Types of derived scores

7.1 *Standard Scores*

These scores use mean and S.D. z score is one type of standard score, expressed as the distance in S.D. units the score differs from the mean.

$$z = \frac{\text{Raw scores} - \text{Mean}}{\text{S.D.}} \qquad (7)$$

Example

Tests	Students	Score	Z score
Test I	A	45	$\frac{45-30}{5}=3.0$
Mean=30	B	30	$\frac{30-30}{5}=0.0$
S.D.=5	C	25	$\frac{25-30}{5}=-1.0$
Test II	A	50	$\frac{50-40}{10}=1.0$
Mean=40	B	40	$\frac{40-40}{10}=0.0$
S.D.=10	C	35	$\frac{35-40}{10}=-0.5$

Thus student A performed much better (3.0) on Test I than on Test II (1.0). The performance of student B is same on both while student C fairs a little better on Test II (−.5) than on Test I (−1.0). Z scores are generally converted by teachers to a scale using whole numbers and avoid negative numbers. One of the commonly used scales is the Z score with a mean of 50 and S.D. of 10 which can be obtained from Z scores by the formula,

$$Z=10Z+50$$

For our preceding example the Z scores are as follows:

(A) $Z=10\times 3.0+50=80$
(B) $Z=10\times 0\ \ +50=50$
(C) $Z=10\times -1+50=40$

(A) $Z=10\times 1\ \ \ +50=60$
(B) $Z=10\times 0\ \ \ +50=50$
(C) $Z=10\times -0.5+50=45$

In both types of computations, the Z scores lead to the same conclusion, differing in their units as under:

(Small)	z score	Mean= 0	S.D.= 1
(Capital)	Z score	Mean=50	S.D.=10

7.2 *Percentile Rank*

A percentile rank of a student indicates the percentage of students whose scores are lower than his. If 80% of students score lower than Amita then Amita's percentile rank is 80. Percentile rank (PR) can be calculated by dividing the number of lower scores by the total number in the class and multiplying the result by 100.

$$PR = \frac{\text{Number of lower scores}}{\text{Total number in the class}} \times 100 \qquad (8)$$

$N = 40$

Example

Score	No. of lower scores	PR
45	39	$\frac{39}{40} \times 100 = 98$
40	20	$\frac{20}{40} \times 100 = 50$
35	10	$\frac{10}{40} \times 100 = 25$
30	4	$\frac{4}{40} \times 100 = 10$

Students can understand percentile rank easily and teachers can report test results meaningfully. Standard scores are more useful for recording unit-wise test results or semester-wise results. In fact both methods are complementary. Standard scores describe how far a student's score is from the average while the PR describe his relative position in the class.

Median is the 50th PR because 50% of scores fall below it. When N is small, the definition of PR may have to be refined. Percentile rank is the score in the distribution below which a given percentage of scores fall "plus one half the percentage of space occupied by the given score". This can be understood by the example given below:

Scores: 60, 67, *53*, 49, 40

In this case 53 is apparently the median i.e. it occupies 50th PR. Thus 50% of scores should fall below it but only two (49 and 40) out of 5 fall below it. This indicates that 53 has a PR of 40 (not 50) because 2 out of 5 scores fall below it. However if we consider the addition in the definition, "plus one half the percentage of space occupied by the given score" we can reconcile as explained below:

Each score (5 in all) occupies 20% of total space.
One half of the percentage (20%) = 10%.
Below 53 lie 40% of the (2 out of 5).
40% + 10% + 0% which is the true percentile rank.

When N is large this qualification is unimportant because the percentile ranks are rounded to the nearest whole number ranging from PR of 99 to that of zero. Percentile rank can be computed by the formula:

$$PR = 100 - \frac{(100\ R - 50)}{N} \qquad (9)$$

where R is the rank from the top.

Example

Roomi ranks 37 in his 10th class of 150 students. Therefore, 36 ranks fall above him and 113 below him. His percentile rank will be

$$PR = 100 - \frac{(3700 - 50)}{150} = 100 - 24 = 76 \qquad (10)$$

7.3. *Percentile Rank and Percentile Score*

A percentile is a score point in the score distribution below which the stated percentage of all measures lies. A student who scores 30th percentile of his class has done better than 30 per cent of students and is poorer than 70 per cent. It is not to be confused with the percentage of the items or questions correctly answered. Percentile rank on the other hand is a point in the distribution below which a given percentage of scores fall. If the 80th percentile rank is a score of 65 then it means that 80 per cent of the scores fall below 65. Likewise in the first example, PR of student is 30.

8. Measures of Relationships

8.1. *Correlation*

Relationships between age and size, intelligence quotient and achievement, height and weight, car size and fuel consumption, reading test scores and tests on literature and such other umpteen examples are quite common in everyday life. This type of relationship has important applications in measurement. The concept correlation indicates the relationship between two or more paired variables or between two or more sets of data. The degree of relationship is measured and presented by the coefficient of correlation. Measures of correlation are therefore:

(a) numerical indexes that reflect quantitatively the extent of relationship between two sets of scores or measures, and

(b) the extent to which examinees hold the same relative position on the two sets of measures.

When the relation is perfect and direct, numerical index is 1.00.

When there is no relationship existing, the numerical index is 0.00.

When relationship is perfect but inverse the numerical index is −1.00.

Most of the measures of interest in the classroom will be between 0.00 and 1.00.

(−1.00)	(0.00)	(+1.00)
High in one trait low in the other	Pure chance relationship. Lack of correlation	High in one trait, high in the other
(a) Age of automobile and sale value	(a) Body weight and intelligence	(a) Intelligence and academic achievement
(b) Time spent in practice and number of spelling errors	(b) Achievement and height	(b) Height and shoe size
(c) Total wheat production and price per quintal		(c) Interest and achievement
		(d) Productivity per acre and value of farmland

Content	*Operations*	*Products*
1. Figural	1. Cognition	1. Units
2. Symbolic	2. Memory	2. Classes
3. Semantic	3. Divergent thinking	3. Relations
4. Behavioural	4. Convergent thinking	4. Systems
	5. Evaluation	5. Transformations
		6. Implications

Therefore $4 \times 5 \times 6$, makes 120 cells or dimensions, each representing a mental process or an ability. Out of these more than 100 abilities have already been identified.

8. Cognitive Domain Taxonomies

(i) *Bloom's Taxonomy* (1956)

The first ever attempt to build a taxonomy was made by a group of college examiners who were confronted with assessment problems. They recognised the need to classify educational objectives into three categories, called the domains. The cognitive domain included recall and recognition of knowledge and development of intellectual skills. The 'affective' domain included objectives relating to interest, attitudes and values while the psychomotor domain includes observed voluntary action or action patterns.

Bloom's taxonomy (1956) is based on four basic postulates (8):

(a) Behaviours designated in the taxonomy are cognitive.
(b) Behaviours are hierarchical in nature.
(c) Behaviours are cumulative in nature.
(d) Behaviours are learned behaviours.

Taxonomy is arranged into six objectives ranging from the simplest to the most complex. Each objective is further defined into sub-categories followed by illustrative items. Taxonomy is as under:

1.0 Knowledge	1.1 Knowledge of specifics
	1.2 Knowledge of ways and means dealing with specifics

	1.3 Knowledge of the universals and abstractions in a field
2.0 Comprehension	2.1 Translation
	2.2 Interpretation
	2.3 Extrapolation
3.0 Application	3.1 Ability to use a theory, principle or a method to solve a problem involving a new or unfamiliar situation
4.0 Analysis	4.1 Analysis of elements
	4.2 Analysis of relationships
	4.3 Analysis of organisational principles
5.0 Synthesis	5.1 Production of unique communication
	5.2 Production of a plan or proposed set of operations
	5.3 Derivation of a set of abstract relations
6.0 Evaluation	6.1 Judgements in terms of internal criteria
	6.2 Judgements in terms of external criteria

(ii) *Gagne-Merrills' Taxonomy* (1965)

This taxonomy integrates cognitive, affective and psychomotor domains. It follows the 'push down' principle which emphasises the acquisition of behaviour at the lower level before acquiring it at the higher level. Eight types of hierarchical behaviours are identified, the learning of one being conditioned by the learning of those which are inferior in the structure. From the easiest to the most complex these are as under (9):

1. Signal learning
2. Stimulus response learning
3. Chaining
4. Verbal association
5. Discrimination learning
6. Concept learning
7. Rule learning
8. Problem solving

(iii) *Madaus* (1973)

While working on the linear hierarchical structure of Bloom's taxonomy, Madaus found a branching structure after the level

Rarely do we find perfect correlation, particularly in relation to human traits. In perfect positive correlation, for every unit increase in one variable there is a proportional unit increase in the other.

Like-wise in a perfect negative correlation, for every unit decrease in one variable there is proportional unit decrease in the other. Thus when two sets of scores correlate, the scores co-vary. In reality, coefficient of +1.00 or −1.00 are not encountered in human traits. The sign of coefficient indicates the direction of the relationship and the numerical value is its strength.

Example

We may examine the following scores on four tests to understand the concept of correlation. Figures 1.1, 1.2 and 1.3 represent the data given in Table 2.

TABLE 2

Student	Test A	Test B	Test C	Test D
1	5	16	28	26
2	10	18	26	18
3	15	20	24	24
*4	*20	*22	22	22
5	25	24	20	28
6	35	28	16	16

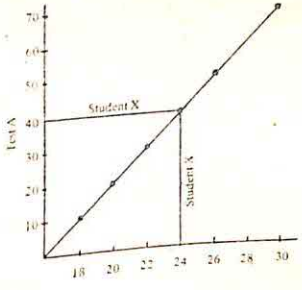


Fig. 1.1

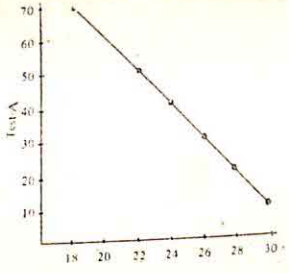


Fig. 1.2

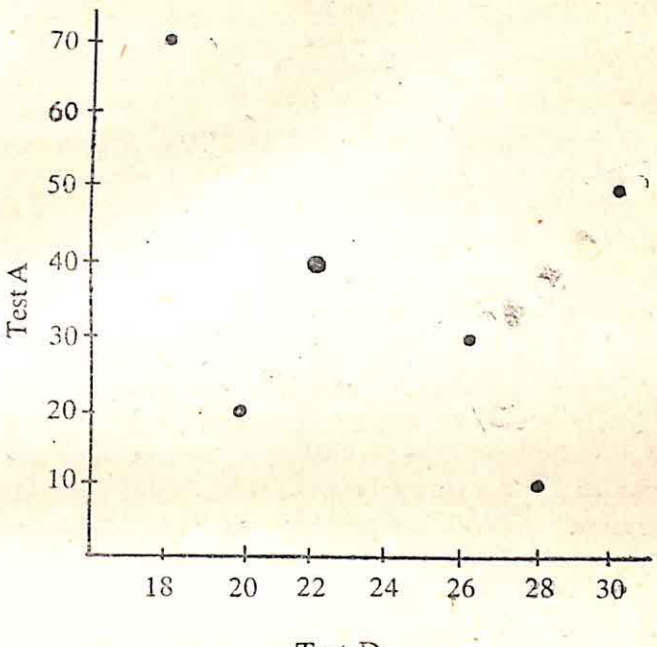


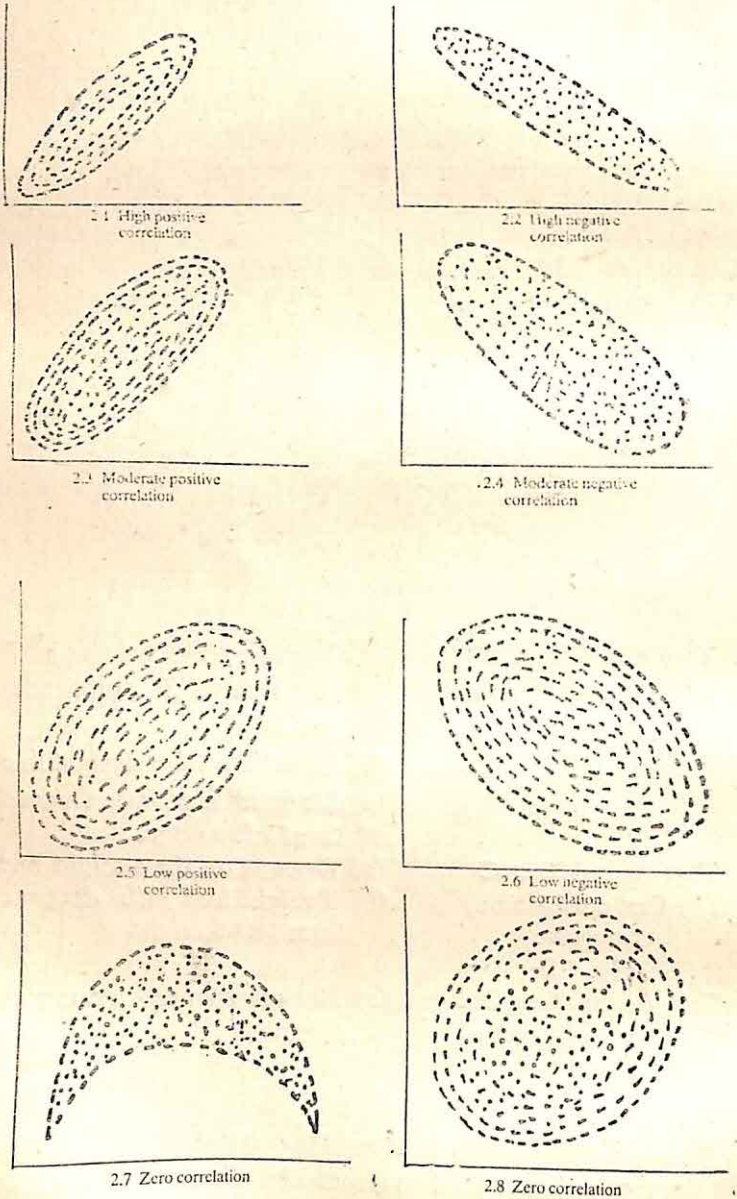
Fig. 1.3

1. Test A and Test B (Perfect correlation) Positive Fig. 1.1	(a) For every 5-point rise in Test A there is 2 point rise in test B (b) Dots lie on the straight line (c) Straight line *rises* diagonally from left to right
2. Test A and Test C (Perfect correlation) Negative Fig. 1.2	(a) Dots lie on the straight line (b) Straight line *falls* diagonally from left to right
3. Test A and Test D (Zero correlation) Fig. 1.3	(a) No covariance among the scores

The degree of relationship between two variables is also depicted by developing scatter-plots from the given data. Following scatter-plots illustrate this. (10)

Figure 2:

Scatterplots illustrating various degrees of correlation



2.1 High positive correlation

2.2 High negative correlation

2.3 Moderate positive correlation

2.4 Moderate negative correlation

2.5 Low positive correlation

2.6 Low negative correlation

2.7 Zero correlation

2.8 Zero correlation

From the above discussion we may understand that:

(a) A correlation coefficient indicates only linear relationship (or lack of it). There can be Zero correlation between two variables that are related in the same way other than the linear fashion.
(b) The coefficient of correlation is only an *estimate* of correlation that actually exists.
(c) Sign (+ or −) has nothing to do with the strength of correlation which indeed is reflected by the size of correlation coefficient. A correlation of +.57 and −.57 are of exactly the same strength, the difference lies in the direction i.e. negative (inverse) or positive (direct).
(d) Correlation between two sets or variables does not imply cause and effect. For example high correlation between the farm output and the consumption of alcohol does not imply cause effect relationship. Increase in economy would cause increase in both.

How to compute correlation estimate?
There are a number of methods to compute correlation.

(i) *Scatter plot method*
The easiest method is to construct scatter plots as shown in Figs. 2.1 to 2.8 and obtain the rough estimate of the size and direction of correlation. Scattered dots indicate low correlation and narrow bands show high correlations.

(ii) *Diedrich (1964) Tetrachoric correlation method* (11)

Steps
1. Rank order the students on the basis of scores on Test 1.
2. Draw a line through the scores at the median.
3. Repeat Steps 1 and 2 for Test 2.
4. Find the total number of students from both the lists, whose names appear in the top half of *both* lists.
5. Divide the number you get in Step 4 by the total number of students taking both tests (i.e. to find percentage of students scoring in the top half on both tests).
6. Locate this percentage (Step 5) in Table 3 given on the next page and note the corresponding correlation coefficient.

TABLE 3: Tetrachoric correlations

Per-cent	r	Per-cent	r	Per-cent	r	Per-cent	r	Per-cent	r
45	.95	37	.69	29	.25	21	—.25	13	—.69
44	.93	36	.65	28	.19	20	—.31	12	—.73
43	.91	35	.60	27	.13	19	—.37	11	—.77
42	.88	34	.55	26	.07	18	—.43	10	—.81
41	.85	33	.49	25	.00	17	—.49	9	—.85
40	.81	32	.43	24	—.07	16	—.55	8	—.88
39	.77	31	.37	23	—.13	15	—.60	7	—.91
38	.73	30	.31	22	—.19	14	—.65	6	—.93

Source: Diedrich, *Short cut Statistic for Teacher made Tests* (pp. 34-36).

(iii) *Pearsons' product moment method* (r)

For most of the classroom work the two methods described above are accurate enough. If a calculator is available Pearsons' product moment correlation can be computed without much effort by using the following formula which is not difficult to use:

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x^2)]\,[N\Sigma y^2 - (\Sigma y^2)]}} \qquad (12)$$

Although the formula appears quite cumbersome, it is not so as you need only six kinds of information. These are:

1. Σx i.e. the sum of raw scores on Test 1.
2. Σx^2 i.e. the sum of the squared scores on Test 1.
3. Σy i.e. the sum of the raw scores on Test 2.
4. Σy^2 i.e. the sum of the squared scores on Test 2.
5. Σxy i.e. the sum of the cross products of the raw scores of Test 1 and Test 2.
6. N i.e. the number of students (not the scores).

(iv) *Spearman's rank order method* (ρ)

Spearman rank order coefficient of correlation is also simple to compute by using the formula:

$$\rho\,(\text{Rho}) = 1 - \frac{6\Sigma D^2}{N(N^2-1)} \qquad (13)$$

where D = the difference between paired ranks

ΣD^2 = the sum of squared differences between ranks

N = number of paired ranks

Example

Two judges Rao and Kamla ranked 10 pupils in a debate contest as under:

To what extent are their judgements in agreement?

Pupil	Rao's ranking	Kamla's ranking	D	D^2
A	5	4	1	1
B	2	1	1	1
C	1	3	2	4
D	10	7	3	9
E	8	10	2	4
F	9	8	1	1
G	6	5	1	1
H	4	6	2	4
I	7	9	2	4
J	3	2	1	1
N=10				$\Sigma D^2=30$

$$\rho=1-\frac{-6\,(30)}{10\,(100-1)}=1-\frac{180}{10\,(99)}$$

$$\rho=1-\frac{180}{990}=1-\frac{2}{11}=.82$$

This means high degree of agreement between the two judges.

(v) *Wolfes' (adaptation) method* (14)

When a teacher is interested in finding the correlation between the scores of students on two unit tests it can be calculated as under. Examine the example given below showing scores of 20 students on unit Test 1 and Test 2 showing marks out of 25.

TABLE 4

Pupil	Scores		Classification	
	Unit test 1	Unit test 2	Unit test 1	Unit test 2
1	2	3	4	5
A	*24	*20	H	H
B	20	11	—	—
C	10	13	—	—
D	7*	14	L	—

(*Contd.*)

1	2	3	4	5
E	15	15	—	—
F	*22	*17	H	H
G	13	10	—	—
H	14	*19	—	H
I	7*	5*	L	L
J	17	*16	—	H
K	*23	16	H	—
L	19	13	—	—
M	12	11	—	—
N	*21	*18	H	H
O	6*	8*	L	L
P	9*	*17	L	H
Q	5*	8*	L	L
R	17	16	—	—
S	*24	6*	H	L
T	18	8*	—	L

How to proceed?

Step 1: Decide how many students to include in high and low group. In usual classes of 20 to 30 it is alright to have one fourth i.e. 5 scores in this example; (21, 22, 23, 24, 24) and (5, 6, 7, 7, 9) in Test 1 and (17, 17, 18, 19, 20) and (5, 6, 7, 8, 8) in Test 2.

Step 2: Cross mark on the left the five high and on the right the five low scores in each test. For example high scores, 24, 24, 23, 22, 21 and low scores 5, 6, 7, 7, 9 in Test I and high scores, 20, 19, 18, 17, 17 and low scores, 5, 6, 8, 8, 8 in Test II are identified.

Step 3: Under columns 4 and 5 write the letter H for the scores that are among the highest 5 on each test i.e. against the cross marks put (on the left) under respective columns.

Step 4: Under columns 4 and 5 write the letter L for the scores that are among the lowest five on each unit test, i.e. against the cross marks (put on the right) under respective columns.

Step 5: Prepare a table like the following indicating the position by tallies or symbols, of H and L in each of the four squares.

		Test-2 L	Test-2 H
TEST-1	H	HL 1 (1)	H H 3 (3)
	L	LL 3 (3)	LH 1 (1)

In this checker board 3 tallies in HH are of students A, F and N.

In LL three tallies are of 1, O and Q. In HL it is of S and in LH it is of P.

Step 6: Calculate the coefficient (r) of correlation by the formula: $r=\frac{(\text{Sum of HH and LL})-(\text{Sum of HL and LH})}{2n}$ (15) when n is the number of H or L group (5 in this case) i.e. $\frac{1}{4}$ of 20.

$$\text{Thus } r=\frac{(3+3)-(1+1)}{2\times 5}=\frac{6-2}{10}=\frac{4}{10}=0.4$$

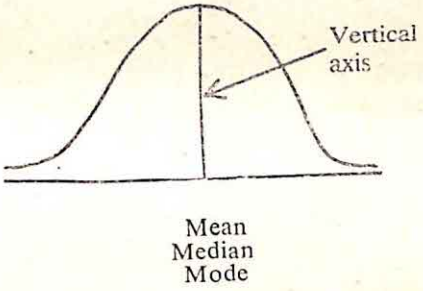
Understandings

1. Coefficient of correlation show the extent to which high and low scores on the test are associated with second test. The closer the index to $\pm$ 1 the stronger the relationship.
2. When r is more than $\pm$ .90 the relationship is quite strong, when r is between $\pm$ .30 the relationship is quite weak. In typical classes relationship is generally between .4 and .60.
3. Correlation coefficient does not reflect proportion of a perfect relationship ($\pm$ 1.0) i.e. r of .75 is not $\frac{3}{4}$ of $\pm$ 1.0.
4. Correlation coefficients do not show proportion of high performers on the test who will score high on the second.
5. Correlation is only a description of general tendencies e.g. when r=.60 between Test I and Test II it indicates that there

is tendency for students who score high on one test to score high on the other.

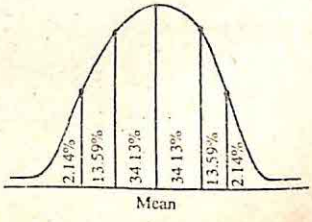
9. Normal Probability Curve

It is a bell shaped curve which is unimodal and symmetrical. It is unimodal because scores cluster round a single point. It is symmetrical because if the left half portion is folded over the right half portion at the mean there would be an exact fit or the mirror image of each other, as shown in the diagram below. Do the "tails" touch the base line? Does it continue to approach more closely as one moves further from the mean?



A normal curve represents the way many human characteristics are distributed. A normal curve is a mathematical model. Since this curve is precisely defined, we can compute the scores falling under any area of the curve. We have already learnt that in a normal distribution, mean, median and mode lie at the same point i.e. they are identical in values.

If we divide the curve by marking to the left and right the units of one standard deviation the normal distribution spans approximately 3 standard deviation on either side. There are very few scores below and above the mean beyond 3 units as seen in the diagram. Curve being symmetrical, area to the

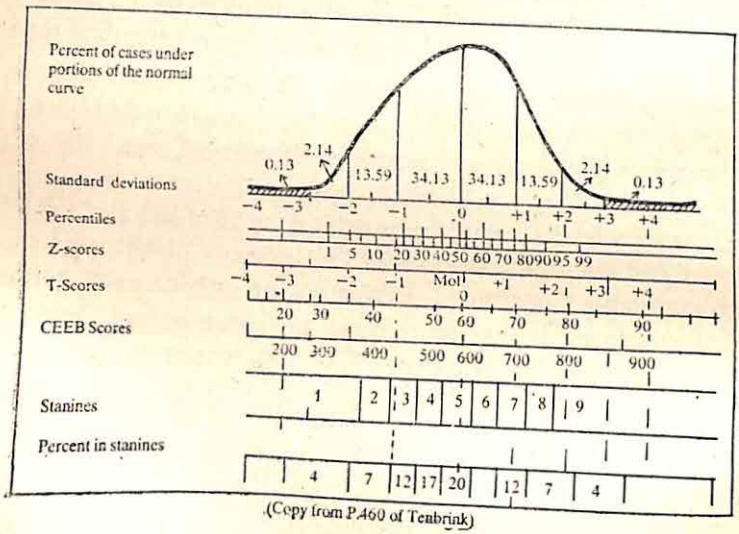


right of the mean is equal to the area left of the mean. Thus the area between mean and 1σ above and below the mean is the same. The height of the curve at a given point indicates the frequency of scores at that point. Therefore, there are as many scores that are one standard deviation above the mean as there are one standard deviation below the mean.

The area of the curve between any two points represents the number of frequency of scores falling between those two points in a normal distribution. The scores falling:

(a) between the mean and 1σ above the mean = 34.13%
(b) between the mean and 1σ below the mean = 34.13%
(c) between mean − 1σ to mean + 1σ = 68.26%
(d) beyond + 1σ and − 1σ cover an area = ⅓ of scores
(e) below the mean − 1σ = 1/6 of the total scores
(f) above the mean + 1σ = 1/6 of the total scores.

This curve is so precisely defined that we can compute the proportion of scores falling under any area of the curve as illustrated in the following figure as depicted by TenBrink (16).



(Copy from P.460 of Tenbrink)

Properties of Normal Curve

1. With the help of equation of the normal probability curve

$$y=\frac{N}{\sqrt{2\pi}}e^{\frac{-x^2}{2\sigma^2}}$$

When N and σ are known it is possible to compute

(a) the frequency (or Y) of a given value x and
(b) the number of percentage between two points or above or below a given point in the distribution.

In this equation:

Y = Scores along the base line or X-axis
y = height of the curve above the X-axis i.e. frequency of a given x-value
N = Number of cases
σ = Standard deviation
π = 3.1416
e = 2.7183
} Constants

2. Total area under the curve is taken arbitrarily to be of 10,000 cases. The curve may be taken to end at points -3σ and $+3\sigma$ distant from the mean, and covers 9973 cases or 99.73% of the entire distribution and disregards .27 or 1% of distribution.
3. Mean, median and mode—all fall exactly at the mid-point and are numerically equal.

10. Reliability

The accuracy or precision with which a test measures whatever it measures is called reliability. There is no test which may be completely reliable (1.0). There is always some error of measurement. So the true score is always contaminated with the error component. The observed score, therefore, consists of true score and error component.

$$X=T+E$$

where X = raw score
E = error score (+ve or −ve)

T and E are never known for any student. They are inferred theoretically on the basis of test taking behaviour.

Error of measurement is the result of the use of imperfect instruments of evaluation and not the mistakes that creep in as a result of handling test scores. Errors of measurement may be due to:

(a) sample of items used in the test,
(b) variation in physical health, and external conditions of the examinee,
(c) chance or luck in guessing,
(d) scoring in case of essay type questions.

Therefore, the test, the conditions and the examinee are the three primary sources of the errors that creep in. Students' score differ due to individual differences in their traits which the test measures and also due to errors of measurement. The degree to which scores differ are described by S.D. of the test but square of the S.D. i.e. *variance* is more useful to depict the relative contribution of the true scores and error scores to the apparant differences among students' performances on a test.

Variance of observed scores can be defined as the sum of the variance of the true score and the variance of error scores as expressed below:

$$V_0 = V_t + V_e$$

where V_0 = variance of observed scores
V_t = variance of true score
V_e = variance of error scores

V_e will be zero on an ideal test and so no error of measurement. Since reliability coefficient is an indicator of the accuracy of scores on a test, we may define coefficient of reliability as:

$$r_{tt} = \frac{V_t}{V_0}$$

where r_{tt} = coefficient of reliability
V_0 = variance of observed scores (S.D.^{2})
V_t = variance of error scores

From this we can note that coefficient of reliability is the

proportion of observed score variance which is due to true score variance. The higher this proportion, the more reliable the test would be. However, reliability coefficient cannot be computed from the above formula.

10.1 *Methods of Computing Reliability Coefficients*

Following methods can be applied by the classroom teachers:

(i) *Split-half*

(a) A test is divided into two halves, usually odd numbered items versus even numbered items.
(b) The two halves are scored separately for each student.
(c) The correlation coefficient is computed between the scores on the two halves.
(d) The estimate of reliability coefficient is obtained by using the Spearman-Brown Formula:

$$r_{tt}=\frac{2r_{\frac{1}{2}\frac{1}{2}}}{1+r_{\frac{1}{2}\frac{1}{2}}} \qquad (16)$$

where r_{tt}=reliability coefficient of the total test

$r_{\frac{1}{2}\frac{1}{2}}$=coefficient of correlation between two half-length tests

Example

Suppose your split-half correlation is .70. Its reliability coefficient will be $=\frac{2\times.70}{1+.70}=\frac{1.40}{1.70}=.82$

Split-half reliability can be used for essay type tests also by obtaining separate scores on odd and even questions e.g. 1, 3, 5 and 2, 4, 6.

(ii) *Test-retest*

(a) A test is given and its scores are recorded.
(b) Same test is given again to the same students at another time with a gap of a few days.
(c) Test-re-test reliability is calculated by correlating scores on first testing and the same students' scores on second testing.

(iii) *Kuder-Richardson formula-20*

In case of objective type questions, where each item is right

or wrong i.e. 1 or 0 mark, this formula can be used to estimate the internal consistency. The formula used is:

$$r_{tt}=\left(\frac{n}{(n-1)}\right)\left(1-\frac{\Sigma pq}{t^2}\right) \qquad (17)$$

where r_{tt} = reliability of the total test
n = number of items in the test
t = S.D. of the test scores
p = proportion of the group answering a test item correctly
q = proportion of the group answering a test item incorrectly (1 − p).

(iv) *Simplified Kuder-Richardson formula-21*

$$r_{tt}=1-\frac{\bar{X}(K-\bar{X})}{K(SD^2)}$$

where $\bar{X}$ = mean
K = number of items in the test
SD^2 = variance (standard deviation squared)

Example

If $\bar{X}=40$, S.D. = 5, K = 60

$$r_{11}=1-\frac{40(60-40)}{60(5)^2}=1-\frac{40\times 20}{60\times 25}$$

$$=1-\frac{800}{1500}=1-\frac{8}{15}=\frac{7}{15}=.47$$

(v) *Saupe method*

J.L. Saupe has given the following formula which is even simpler for approximating Kuder-Richardson results.

$$r_{tt}=-\frac{.19K}{(S.D.)^2} \qquad (18)$$

where r_{tt} = reliability coefficient
SD^2 = variance
K = number of items in the test
.19 = constant

Example

Suppose K=60, S.D.=5

$$r_{tt}=1-\frac{0.19\times 60}{25}=1-\frac{11.40}{25}=1-.45=.55$$

Since different methods involve slightly different assumptions about the nature of data and reliability there would be some variation in the results obtained from different methods. Saupe's formula yields lower estimates than the Split-half method although the two results do not differ appreciably.

Saupe's formula is neither suitable for tests which are not scored on 1−0 basis for correct and incorrect answer nor for those which are very easy or very difficult. Split-half method is suitable for those tests which assign equal number of marks or points for each correct answer as in case of multiple choice tests and attitude scales. Neither of the two methods is suited for essay tests and where speed is a primary factor.

Reliability is only an *estimate* and should be interpreted quite cautiously. It refers to the group tested. The same test if used for a different group will have different connotation as far as interpretation is concerned. The reliability coefficient of a test will be much higher with a group whose mean score is 75 as compared to the one with mean score of 35. Factors affecting reliability in case of objective type, short answer and essay type will be discussed in more detail in the chapter dealing with the qualities of a good test.

11. Standard Error of Measurement

Reliability and standard error of measurement are inter-related concepts. Reliability indicates freedom from error while standard error of measurement estimates the amount of error that *does* exist. While reliability is used as an estimate of the accuracy of the results of measurement as a whole, the standard error of measurement is used as an estimate of the accuracy of the result for a given individual. Measurement error occurs when the observed score of a student differs from his true scores as we learnt earlier X=T+E. Variance of the observed score

(V_0) is the sum of variance of true score (V_t) and variance of error (V_e)

Thus $V_0 = V_t + V_e$.

For any one student, true score and error score cannot be known directly but standard deviation of the error scores can be estimated on the basis of which individual performance can be interpreted. This standard deviation of the error of measurement is called the *standard error of measurement* (SEM). It can be estimated by using the following formula:

(i) $SEM = SD\sqrt{1-r_{tt}}$ (19)

where SEM = standard error of measurement (also called error of an obtained score)

SD = standard deviation of the test scores

r_{tt} = reliability coefficient of the test

Example

In our earlier example SD = 5 and $r_{tt} = 0.55$

$$SEM = SD\sqrt{1-r_{tt}}$$

or $= 5\sqrt{1-.55} = 5\sqrt{.45} = 5 \times .67 = 3.35$

This SEM of 3.5 means that standard deviation of the errors of measurement on this test is 3.5. This means that the average distance between the observed scores and the true score is estimated to be 3.5. Thus if Anjum's obtained score is 58, what will be her true score? Assuming that we did not commit greater errors in her score, Anjum's true score is no more than 3.5 points (plus-minus) from her observed score of 58. This would mean that her true score lies somewhere between 61.5 and 54.5 ($58 \pm 3.5 = 54.5$ to 61.5).

Lord's Method of Calculating SEM

If the test is neither extremely easy nor extremely difficult and admits only one correct answer for each item thereby allowing for 1 or zero mark for each correct/incorrect response, Lord's method of calculating standard error of measurement can be used and is quite simple.

$SEM = \sqrt{0.432K}$ (20)

where K = number of item

0.432 = a constant

Thus for a test of 60 items

$$SEM = 0.432 \quad \sqrt{60} = .432 \times 7.75 = 3.35$$

This value is almost the same as calculated by the 1st method.

Readymade tables are available to estimate the SEM directly from the number of items (K). A few SEM figures corresponding to item number are given below for ready reference.

Item no. (K)	SEM	Item no. (K)	SEM	Item no. (K)	SEM
10	1.4	45	2.9	75	3.7
15	1.7	50	3.1	80	3.9
20	1.9	55	3.2	85	4.0
25	2.2	60	3.4	90	4.1
30	2.6	65	3.5	95	4.2
40	2.7	70	3.6	100	4.3

Errors of measurement are normally distributed according to test theory. On the basis of this theory we can predict that a student's observed score will be no more than

(a) two standard errors of measurement higher or lower than his true scores in 95% of the cases
(b) one standard error greater or less than his true score in 68% of the cases.

Thus standard error of measurement can be used to establish confidence limits for students' score within which a student's true score is likely to fall. In our previous example used to explain SEM of the test, the 68% confidence limit is 38 ± 3.5 i.e. 54.5 lower limit and 61.5 upper limit. This means that there are 68 chances out of 100 that a student with 48 marks (observed score) on a test will have a true score between 54.5 and 61.5. This concept uncovers the fallibility of test score

about which a test user must be aware for describing a student's status.

For assignment of grades there should be a difference of 3 standard errors of measurement wide for each category. This avoids misclassification of students by more than one grade category. A grade category is generally one standard deviation in width. A classroom test should have at least .89 the reliability coefficient for all the five grade categories normally used (A, B, C, D, F).

12. Standard Error of Estimate

The validity of a test is the extent to which it measures what it intends to measure. It indicates the extent to which intended purpose of a test is served. Whereas reliability indicates how *free* from errors our judgements are. The standard error of estimate goes to estimate the *amount* of error we *do* make while making a judgement. While validity is used as an estimate of the accuracy of making particular judgements, the standard error of estimate is used for estimating the accuracy of a particular judgement for a given individual.

Unlike SEM which indicates how far obtained scores are away from the true score, the standard error of estimate indicates how far off the estimated scores or predicted scores will be from the scores actually obtained. Take for example a scientific aptitude test given at the end of class VIII to use the result for selection of students for science stream in class IX. Suppose the standard error of estimate based on the predictive validity coefficient of the aptitude test is 4.5. We can now interpret like the SEM, the standard error of estimate, that the counsellor is reasonably certain that Anita would score between 85 ± 4.5 i.e. 80.5 to 89.5 marks.

From this discussion of SEM and standard error of estimate it is evident that these two concepts are very useful in estimating the confidence level or bands presenting the lower and upper limits of true scores, thereby providing the needed basis for meaningful interpretation of observed scores that accrue in classroom tests.

13. To Sum Up

In this chapter we studied the basic statistical concepts which are pre-requisite to the understanding of measurement concepts in the forthcoming chapters. How scores obtained by use of classroom tests can be made more meaningful by applying basic simple statistical concepts, are discussed. Organisation of data into frequency distribution, calculation of measures of central tendencies, percentile range, standard deviation, quartile deviation followed by discussion on measures of relationship like correlation are briefly discussed in a non-technical language. Calculation of reliability and standard errors of measurement are discussed along with their implications. A deliberate attempt has been made to clarify these concepts with the help of examples. More simple and usable formulae involving less technical and mathematical know-how which can be applied by an average classroom teacher, are given. The major purpose of including these fundamental concepts before the actual text on measurement and evaluation is that the reader would be able to comprehend better the explanations that follow in forthcoming chapters, and appreciate how simple basic statistical concepts are useful in manipulating data to facilitate meaningful interpretations.

CHAPTER IV

INSTRUCTIONAL OBJECTIVES AND EVALUATION

1. Introduction

More often than not the word objective, goal and aim are used interchangeably in educational literature. This causes not only confusion for the students of evaluation in understanding the process of education but also creates difficulty in fixing time targets for instructional outputs at various levels which become guide posts for teaching-learning sequences and strategies. To what extent has an instructional programme succeeded? What are the indicators of success or failure? For finding a rationale to these questions, one has to look for some reference point or the criterion against which the outcomes of instruction can be judged for its success or failure. This reference point or the criterion is the 'objective', 'goal' or 'aim' for which instruction was planned. It is therefore, necessary that first of all we should be clear about the connotation of these words. Once this is understood, further details about their sources, levels, classification, attributes, types, statement, specification and role in the teaching-learning process can be better appreciated.

As a result of the study of this chapter you should be able to:

(a) recognise the difference in nature and scope of the word's objective, goal and aim,
(b) derive objectives from various sources like learner, society

and knowledge,

(c) identify objectives at various levels from national to classroom levels,

(d) classify objectives into various categories according to different taxonomies,

(e) detect errors in statement of various objectives,

(f) define different objectives in terms of behavioural outcomes,

(g) analyse the implications of criterion-referenced and norm-referenced objectives in evaluation,

(h) establish the relationship between objectives, learning experience and evaluation,

(i) appreciate the role of objectives in teaching, learning and testing.

2. Nature of Objectives

An objective is a normative concept which carries with it the idea of goodness or the desirable. What is desirable, depends on what we consider good for the learner. Who decides what is good and desirable? Society sanctions the desirability or goodness of the behaviour of an individual. This in turn depends on the values that the society cherishes. Thus what sort of individual or learner do we visualise as a result of our having educated him through the instructional programme? This means particular types of behaviours or desirable changes that we can expect our students to acquire after undergoing education. It is these intended changes or learning outcomes or expected modes of behaviours that we like to appraise also in the pupils in order to judge the effectiveness of the educational process.

If society considers manual work and the dignity of labour as a cherished value in a secular democracy, this value is made explicit in the form of objectives that are stated in the course of study as 'development of the dignity of labour among students'. Socially useful productive work then becomes a new discipline or a course of study to achieve that objective. Therefore, the intended product of learning and not the content of learning is reflected in an objective. Although an objective refers to values which are judged as desirable and are given priority for

transmission through the educational process, yet it is not a value by itself. It is a product of value judgement. Likewise, an activity, content, a learning situation or an experience or a process of learning, the medium of learning etc. are all means towards attainment of intended or expected learning outcomes but are themselves not the end product of learning. Thus an objective is concerned with the product of learning rather than with the process of learning. The difference between the process and product of learning can be appreciated better by the following example given in a schematic manner based on Gronlund (1).

Learner	When Undergoes →	Planned Learning Experiences (Process)	Leads to →	Intended Learning outcomes (Product)
		↓		↓
A. Study of Petunia flower in the Laboratory				
		(a) Exposition by teacher		(a) Knowledge of floral parts of Petunia
		(b) Pupil teacher discussion		(b) Ability to describe flower in semi-technical terms
		(c) Pupils' observations		(c) Skill in observing
		(d) Section cutting by students and demonstration by teacher		(d) Skill in section cutting
		(e) Drawing diagrams by students		(e) Skill in drawing diagrams of flower

From this illustration it is obvious that

(i) objectives indicate direction of students' growth (a—c).

(ii) objectives vary in their complexity; a is simpler than b.

(iii) certain objectives (c—e) can be developed through varied content (not only Petunia flower) while others (a) have specific content.

(iv) certain activities (Process) lead to development of more than one objectives e.g. observation leads towards objectives a, b and c.

(v) more than one activity may lead towards the same objectives e.g. activities a, b, c and e lead towards objective a.

Therefore, an objective represents the end point towards which activities are directed. They reflect the purposefulness of the teaching-learning process.

To sum up

An objective is *not*

1. a life value but a product of value judgement.
2. a process but a product of learning.
3. an activity but a learning outcome.
4. a method of teaching but expected outcomes of instruction.
5. content of learning but an outcome of learning.

2.1. *Goals and Objectives*

Both these terms are often used synonymously. However, educational planners and evaluators use the term 'goal' as a general statement of desired outcomes, aims, or purposes having long-range implications and involving complex human behaviours. They are useful in stating the purposes of education, purposes of a curriculum to express a view point, in identifying priorities for a policy statement and to communicate with the layman. Statement of goals helps the teachers in communicating programme goals to administrators, students and parents besides conceptualising the desired outcomes in proper perspective.

Following are the goal statements:

(a) To develop a wholesome personality.
(b) To develop physical fitness.
(c) To develop a literate citizenry.
(d) To develop ethical and spiritual values.

Since these statements are global it is difficult to know from direct observations whether these goals are achieved.

An aim is also a general statement which reflects a specific focus, a level of attainment or a point towards which action is directed. Goals are ultimate aims while specific aims may be called objectives. Likewise general objectives are aims. Aims

are long-ranged while objectives are short-ranged when compared to each other. Even objectives (instructional) may be terminal and short term, depending upon the time required to achieve those objectives. Therefore, to reach a goal one may aim at a particular level followed by another level till one reaches the goal. In order to achieve the pre-determined aim one may have proximate aims or objectives, the attainment of which becomes pre-requisite to achieve the aims. Thus objectives are specific aims with more precisely defined targets or aims achievable in an observable time limit. These objectives can be further classified as terminal and short-term instructional objectives. The following examples will clarify further the relationship among goals, aims and objectives.

Goal : To develop a literate citizenry for Indian democracy.
Aims : To universalise primary education.
Terminal Objective : By the end of the year 1980, 20% of the students will be able to read and write simple sentences in their mother tongue.
Short-term Objective : Students will write their name correctly at the end of two weeks.

From this relationship we can infer that

(a) Objectives, aims and goals reflect the end product of learning at various levels of attainment.
(b) There is hierarchy and complexity in the behavioural outcomes as we proceed from short-ranged objectives to long-ranged objectives, aims and goals.
(c) The higher the level of statement the more global the objective becomes and the more difficult to identify observable outcomes.
(d) The more comprehensive or global is the statement, the more areas of disciplines it encompasses for its attainability.

2.2. *Objective as Ends and Means*

From the relationship between the three terms it is obvious that whatever the level of the product of learning or learning outcome may be, a hierarchical sequence is present in their attain-

ment. In other words, lower level outcomes are achieved first before proceeding to the next higher level. This means that an objective at Level 1 becomes a means to the attainment of an objective at Level 2 which in turn becomes a means to the attainment of objective at Level 3. Therefore, objectives indicate intended learning outcomes at different levels, thereby providing direction to pupils' growth. They provide a basis for planning and organisation of the learning experience and selection of an evaluation instrument. It is through objectives that a link is established between teachers, evaluators, parents and students by focussing their attention on the intended product of learning.

3. Sources and Derivation of Objectives

The sources of objectives can be traced to various empirical approaches or to the significance of behaviour considered desirable in a particular social milieu. Both these approaches are discussed below.

3.1. *Empirical Approach*

Whenever we think in terms of the desirability of the objectives, we are looking for behaviours which are desirable or acceptable. Which behaviours should be expected from a student or a worker who has undergone particular course of instruction? (2) Furst refers to the various approaches like job analysis, critical-incidence-technique, diagnostic studies etc., are used for this purpose. In job analysis, those activities are identified that help the performer to do the job well. These activities viz-a-viz the behaviours become the objectives of instruction. Likewise in the critical-incidence-technique, the significant or critical behaviours are noted during the course of the activities undertaken by the performer. Such outstanding behaviours are supposed to be desirable. It is assumed that when such behaviours are acquired by the learner, they would enable him to do his task well. Such critical behaviours are chosen as the intended behaviours or the expected learning outcomes. In case of diagnostic studies, the behaviours of good performers are compared with those of poor performers on the basis of which the list of those behavi-

ours that determine good performance on the job is drawn out. This list then becomes the intended behaviours, the basis for instructional programmes.

Obviously, all the above mentioned methods depend upon certain tools and techniques which become our guide posts. This approach makes an evaluator the servant while the tools and techniques act as his master. In fact, it is not the type of activity or the tool but it is the significance of the activities that should matter. Which activity is desirable, involves a question of values. Therefore, the choice of activities depends on what is considered good and what is desirable. Which activity is good and acceptable is determined by society. Activities like thieving are not considered socially desirable and therefore, cannot be taken as one of the objectives in the process of learning. Activities which promote intended behaviours representing preferred values, determine the choice of the desirable and are made explicit in the form of educational objectives.

3.2. *Study of Different Sources*

Although the formulation of objectives depends on value judgements they cannot wisely be obtained by individuals or committees making such judgements off the cuff or by accepting long-standing practices by critically examining different sources. Three fundamental sources of deriving objectives are the learner, the society and the fund of knowledge in different subjects.

(a) *Needs of the learner*

The most important source of objective is the information about students. What are their needs and abilities? What are the understandings, habits, attitudes and interests that the students need to develop? How do students learn better? What are the things that they can or they cannot learn at a particular maturity level? Such factors have to be considered while determining the end product of learning. The efforts are to be made to impart knowledge, develop understandings, and cultivate attitudes and habits and to capitalise the basic interests. Thus a careful study of the learner and his needs provides good information for deriving objectives of the curriculum.

(b) *Demands of society*

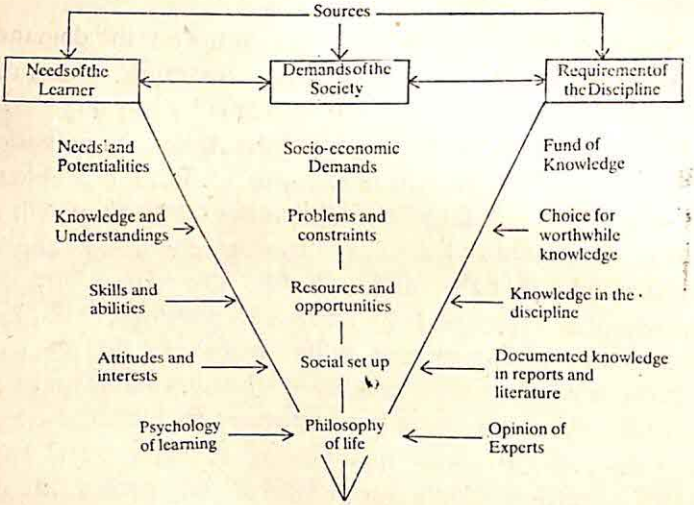
Investigation of contemporary life indicates the demands and commands of the society. It reflects constraints and problems which the young pupils have to overcome by meeting the changing and challenging conditions of the time: the activities they will be expected to perform in the near future; the problems that they are likely to face; the difficulties that they will run up against; opportunities for self-realization etc. Much depends on the philosophy that the nation holds. The nation with a democratic social set-up expects its citizens to develop critical mindedness and the decision-making skills. What are the financial and manpower resources that condition the attainability of certain objectives? What are the other important factors that put constraints on qualitative and quantitative expansion and improvement of education? Such an analysis of our present life reveals the kind of behaviours that students should develop in order to become better citizens of tomorrow. Therefore, societal needs and social philosophy are basic to the derivation of objectives.

(c) *Requirement of discipline*

The nature and developmental level of knowledge in a discipline is also useful in determining the educational objectives. Since vast knowledge is accumulating at a tremendous rate we have to make a choice between what is worth-teaching or worth-learning and what is worth-eliminating. How best can the nature of the discipline be reflected with ease and economy? What is the opinion of experts and specialists in the various disciplines? What do the reports of various commissions and committees say about the preferred objectives? Answers to questions like these will go a long way in determining the educational objectives. (See chart on the next page).

3.3. *Criteria for Selection*

If the list of specific objectives is prepared it may run into hundreds in a particular subject. It is, therefore, essential that selection must be made on the basis of some criteria agreed upon by the specialists in the field. Following criteria may be



useful for selection of objectives:

(i) Objectives should be worthwhile and have educational significance.

(ii) They should agree with the broader goals of education.

(iii) They should be in accordance with the psychology of learning.

(iv) They should be comprehensive enough to cover all areas of human development.

(v) They should be attainable under the school conditions.

(vi) They should be testable in terms of observable and varifiable changes.

(vii) They should be acceptable to teachers from our stand point of teaching resources, availability of time and instruction.

4. Stating Objectives at Various Levels

Objectives can be identified and stated at various levels right from the national level to the periodwise or lesson level. National level objectives emerge from the constitution which provides the values to the society. The objectives at the national level are most global. The Kothari Commission (3) lists four major objectives, viz. education for production, education for national integration, education for modernisation and education for development of spiritual and moral values. It is these objectives which become the source for formulating objectives

at the state level, which in turn pave the way for delineating stage-wise objectives (elementary, secondary and university). Then comes the objectives at the subject level which are drawn in conformity with the stage-wise objectives. From the instructional objectives of a subject we identify the objectives of a unit of teaching. Unit-wise objectives are basic to listing of period-wise or lesson-wise objectives.

Thus we see that a link is established among objectives at various levels. There is some invisible thread connecting the objectives between any two levels. At the highest level the objectives are too abstract to be achievable but as we descend from a higher level to a lower level, they become more concrete. Who are the persons to formulate objectives at various levels? From the student and the classroom teacher to economists and politicians, are involved depending upon the stage at which these are to be formulated. The following chart indicates the personnel who are usually considered appropriate. In actual practice, however, all state agencies concerned with curriculum development do not always take advantage of the expertise indicated in the chart given below:

Level	*Suggested Personnel*
(a) National level	Educationists, Economists, Sociologists, Politicians, Scientists, Psychologists and Administrators
↓	
(b) State level	Educationists, Administrators, Economists, Sociologists, Curriculum Experts and Teachers
↓	
(c) Stage level	Educationists, Administrators, Curriculum experts, Teachers, Parent Associations
↓	
(d) Subject level	Subject experts, Curriculum experts, Teachers and Subject associations
↓	
(e) Unit level	Curriculum experts and Teachers
↓	
(f) Lesson level	Teachers and Pupils

To appreciate the link between the objectives at various levels we may examine the following illustration of objectives stated at all the six levels.

L-1. To modernise education for best exploitation of natural

7. Classification of Educational Objectives

Depending upon the basis of classification, educational objectives can be categorised as under:

7.1. *Two-way Horizontal Classification*

Criterion	Classification	Illustrations
1. Level of generality	(a) General versus specific	(i) Ability to solve problems (ii) Ability to add whole numbers
2. Time contingency	(b) Ultimate vs. proximate	(i) To develop good health habits (ii) To know four basic food groups
3. Concreteness or tangibility	(c) Tangible vs. intangible	(i) To acquire knowledge of social studies facts (ii) To develop social sensitivity
4. Course coverage	(d) Single vs. multiple course objectives	(i) To develop dissectional skills among students (ii) To develop ability to judge
5. Operativeness	(e) Stated vs. functional	(i) To develop critical mindedness (ii) To identify facts from opinion OR to identify underlying assumptions
6. Area of development	(f) Cognitive vs. non-cognitive	(i) To develop understanding of parts of speech (ii) To develop appreciation for music (iii) To handle clinical thermometer efficiently
7. Type of response (varied or stipulated)	(g) Open vs. closed	(i) To develop a plan of action for reducing noise pollution (ii) To identify three past perfect sentences in a given para
8. Type of learning outcome	(h) Product vs. process	(i) To develop concept of osmosis (ii) To develop observing power of pupils

Likewise more bases could be used to have two-fold or three-fold classification of educational objectives.

7.2. *Multiway Horizontal Classification*

A number of attempts have been made to classify objectives. As early as 1918 the Commission on Reorganisation of Secondary Education in America gave the following objectives (5):

(a) Health
(b) Command of the fundamental processes
(c) Worthy home membership
(d) Vocation
(e) Citizenship
(f) Worthy use of leisure
(g) Ethical character

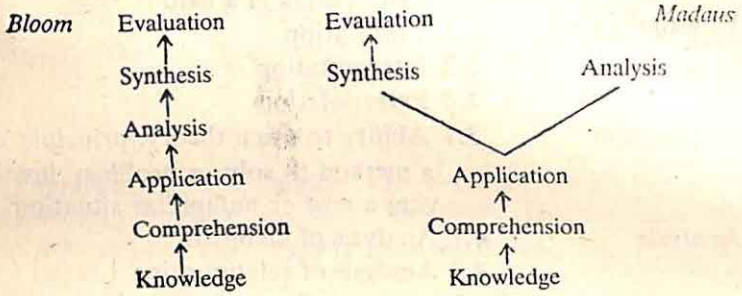
7.3. *Multiway Vertical Classification*

(i) *Ebel* (1966) classified the various classroom tests into following categories (6):

(a) Understanding of terminology	(Vocabulary)
(b) Understanding of facts and principles	(Generalisations)
(c) Ability to explain or illustrate	(Relationships)
(d) Ability to calculate	(Numerical problems)
(e) Ability to predict	(What is likely to happen under specified conditions)
(f) Ability to recommend appropriate action	(In some specific practical problem situation)
(g) Ability to make an evaluative judgement	

(ii) *Guilford* (1967) in his structure of intellect has described three faces of intellect, namely the content, operation and the product each of which is defined as under (7):

of application as shown below:



(iv) *Decorte* (1973)

The four dimensions of this classification model are:

1. The subject matter of specific content of a given universe of objectives.
2. The domain of information to which the subject matter belongs.
3. The product: The formal aspect of information the objectives produce.
4. The operation: as defined in Guilford's model:

A. Receiving-reproducing
   - 4.1 Perception of information
   - 4.2 Recall of information
   - 4.3 Reproduction of information

B. Productive operations
   - 4.4 Interpretive production of information
   - 4.5 Convergent production of information
   - 4.6 Evaluative production of information
   - 4.7 Divergent production of information

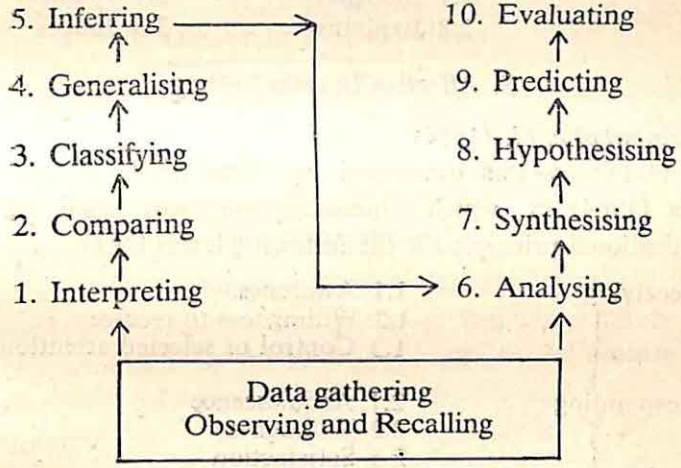
This system integrates Bloom and Guilford's models and becomes a tool for defining cognitive objectives.

(v) *Hannah and Michaelis* (1977)

The framework of instructional objectives given by these authors gives four categories or types of learning, each of which has a series of levels within it (11).

1. Data gathering is the first category which serves as a foundation for the other three types of learning. Observing and recalling are the two bases for data gathering which lead to the other three categories mentioned below:
2. Intellectual processes (Cognitive domain)
3. Skills (Psychomotor domain)
4. Attitudes and values (Affective domain)

Cognitive domain (Intellectual processes)
According to the authors, intellectual processes are all actions involving mental or cognitive abilities arranged in a hierarchical order based on increasing complexity as shown below:



(vi) *Singh* (1977)
Working on empirical validity of Bloom's taxonomy the author himself found that as far as analysis and synthesis are concerned, these categories seem to be redundant as they appear to be a part of the application objective as evidenced from empirical validity. The other categories seem to be hierarchised as shown here (12).

Knowledge → Comprehension → Application → Evaluation

(vii) *N.C.E.R.T. Taxonomy* (1986)
N.C.E.R.T. taxonomy is based on more than 20 years of usage in the field. It is validated by thousands of teachers involved in evalutaion workshops for development of objective based

resources of the country.

L-2. To conserve natural resources by preserving forests.

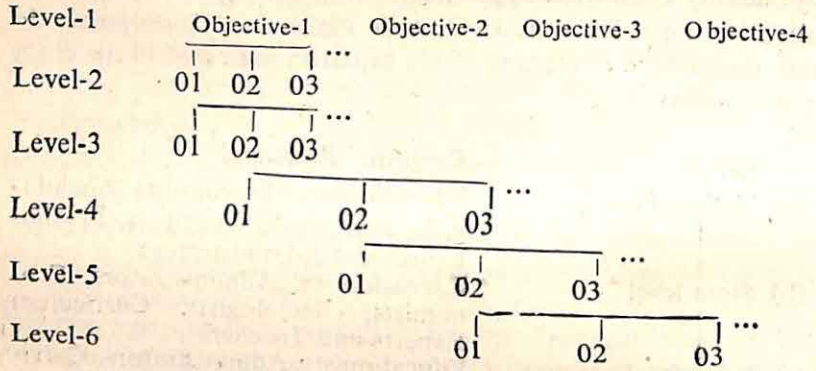
L-3. To conserve animal life for maintaining balance in nature.

L-4. To understand the various food webs, food chains and food pyramids in a forest.

L-5. To study various types of food pyramids.

L-6. To know the concept of producer, consumer and decomposer.

The list is only partial, just to show the nature of objectives which can be traced from one level to the other level downward. In fact at each level there will be more than one objective and each objective when further derived will provide a list of more than one objective at each level as shown in the following symbolic derivation from level 1 to level 6.



Similarly the objective 2, 3 and so on will be worked out downward. The process of deriving objectives can be summarised as follows:

1. Develop understanding of the role of objectives and the need for formulating a list of objectives.
2. Identify and state the general purpose of the programme or the project so that participants in the educational process look to the common referents.
3. Trace the programme or project goals from item-2 above. This may be in operational terms indicating the logical sub-

divisions.

4. Sub-categorise each goal into a statement of objectives with more specific behaviour referents.
5. Develop a time frame in statement of objective so that terminal objectives can be differentiated from enabling objectives.
6. Revise the list of goals and objectives in the light of experience gained in planning, operating and evaluating the programme.

5. Attributes of Usable Objectives

By and large curriculum specialists agree to the statement of objectives in the form of a complete statement. They should also be comprehensive enough to include cognitive, affective and psychomotor areas of students' development. Since the list of objectives may be quite lengthy it may be necessary to curtail this sometimes keeping in view the needs and resources. Therefore the priority-order of goals and objectives must also be specified so that emphasis may be given accordingly.

While stating an instructional objective, the following elements may be noted for understanding the complete statement of an objective (4).

5.1. *Elements of an Objective*

(a) *Content*: What is the object of study? What content, i.e. a fact, concept, principle, or any other content element is the facts of learning. Thus a content object is basic to the statement of any objective. Three basic types of fruits, four types of adjectives, factors affecting climate etc. are examples of object/content.

(b) *Behaviour communicativeness*: Specification of the behavioural elements is essential to indicate that an objective has been achieved. What behaviours should a student exhibit to demonstrate the understanding of 3 types of fruits? Defining the 3 types of fruits or citing illustrations of each type or classifying the given fruits into 3 categories etc. could be an intended outcome, e.g. the ability to classify fruits.

(c) *Conditions of occurrence*: Under what conditions and facilities can the student work and show his mastery of the objective? Conditions should be specified to describe special arrangement.

Example
Give the basis, the students will be able to classify the different fruits into various categories.

(d) *Performance level*: What is the acceptable level of performance? What is the criterion of success? Should students classify fruits with 100% accuracy, 90%, 33% or some other level of accuracy?

Example
Given the basis, students should be able to classify 9 out of 10 fruits correctly.

(e) *Target audience/individuals*: Should 100%, 90% or some other percentage of students achieve the objective?

Example
Given the basis, 80% of the students should be able to classify 8 out of 10 fruits correctly.

(f) *Time of performance*: When should the objective be attained: at the end of the year, term, unit or a lesson? This should be reflected in the objective.

Example
Given the basis, 80% of students should be able to classify 8 out of 10 fruits correctly by the end of the unit.

(g) *Appropriate level of specificity*: Since an objective can be specified at different levels ranging from the most global to the trivial details—like the ability to add two digit number, effort should be made that the list of specific objectives should be so developed that it is neither two global to be tested nor too fragmentary to become unwieldy. The following illustration will clarify this point:

The pupil *understands*

(a) the biological terms facts, concepts, principles etc.
(d) the chemistry of various physiological processes in man.
(c) the mechanism of digestion in man.
(d) the role of various enzymes in digestion.
(e) the function of ptyaline in digestion.

In this example the objective is specified at five levels. Depending upon the level at which you are teaching, an objective can be stated.

5.2. *Terminal, Enabling and Facilitating Objectives*

Terminal performance objective is the desired outcome of learning experiences expressed in terms of the observable behaviour of the learner. According to Magar (1962) "It is a description of a pattern of behaviour we want the learner to be able to demonstrate". (4) Behaviour is overt and is recognisable at the end of a specified time. Output objectives and task descriptions are other terms used synonymously in meaning (5) by writers like Gagne (1965).

Example
To improvise an apparatus illustrating the process of osmosis when the needed equipment is available in the laboratory.

An *enabling objective* is the desired observable behaviour that occurs earlier in time than the terminal behaviour i.e. it becomes an entering or pre-requisite behaviour for the terminal performance behaviour. Just as addition is pre-requisite to multiplication; understanding the concept of osmosis is pre-requisite to improvisation of the experimental set up for demonstrating osmosis, although both are expressed in the same way i.e. in terms of the specific observable behaviour of the learner. The *facilitating objective* generally refers to the behaviour of others (administrators, planners etc.). They involve activities such as making a film, developing learning units, preparing a teacher guide, remedial exercises etc. Such objectives, therefore, facilitate the attainment of terminal and enabl-

ing objectives.

5.3. *Open and Closed Objectives*

Open objectives are descriptive while closed objectives are prescriptive. Therefore, the former involves divergent thinking leading to varied responses while the latter involves convergent thinking calling for one correct or the best response. Some objectives, however may contain both closed and open elements. Both types of objectives are used in specifying learning outcomes. The following examples are representative of the two types.

1. Given 10 out of 50 difficult words, the student will write the correct spellings of 10 words within two minutes. (closed)
2. In an outline diagram of an animal cell the student will draw the four major organelles as given in the text. (closed)
3. After reading the chapter on pollution each student will suggest three measures of his own to check water pollution. (open)
4. Given an unseen passage the students will write its summary in about one third of its length. (open)
5. The student will demonstrate his ability to interpret the given diagram of an experimental set up by

   (a) identifying the three major components of the experiment. (open)
   (b) describing two points relating to control and experimental variables. (open)
   (c) stating the purpose of the experiment. (closed)

The first and second objectives are closed, the third and fourth are open while the fifth is a combination of both.

5.4. *Overt and Covert Behaviour*

Overt behaviour is observable while covert behaviour is not observable. Listing, describing, classifying, stating and answering questions are examples of overt behaviours. Intellectual

processes like interpreting, generalising and evaluating are covert behaviours as also responding, performing and other behaviours reflecting attitudes and values. Covert or primary behaviour is clearly identified as it is persecuted first while overt behaviour or secondary behaviour follows the covert behaviour and indicates a guide line for the assessment of learning.

The following two examples are illustrative—

1. Students will demonstrate their ability to interpret by drawing a floral diagram of a given flower.
   (Interpret is covert behaviour and drawing floral diagram is overt behaviour.)
2. Student will *apply* their punctuational skills by *writing* a paragraph with correct punctuation marks.
   (Apply is covert behaviour and writing is overt behaviour.)

6. Stating Instructional Objectives

Different curricula show different modes of statement of objectives. The following list of objectives together with the improved version indicates the criteria for stating objectives and the common errors observed in statement, of objectives at the senior secondary stage.

Poor	Better
(1)	(2)
(i) To enable the pupils to understand scientific principles.	(i) The pupil understands scientific principles.
(ii) The pupil studies the clinical thermometer thoroughly.	(ii) The pupil handles the clinical thermometer efficiently.
(iii) The pupil understands and applies biological principles to new situations.	(iii) The pupil applies biological principles to new situations.
(iv) The pupil develops skill.	(iv) The pupil develops skills in reading a barometer.
(v) (a) The pupil develops problem solving skills in mathematics. (b) The pupil develops the	(v) The pupil develops the ability to analyse a given problem to know what is given and what is required.

test items, topic tests and sample question papers, organised in collaboration with various examining boards. Objectives under the cognitive domain are listed hereinafter (13).

NCERT Taxonomy Objectives

1. *Knowledge*	2. *Understanding*	3. *Application*
1.1 Recalls	2.1 Translates	3.1 Analyses
1.2 Recognises	2.2 Illustrates	3.2 Hypothesis
	2.3 Identifies	3.3 Suggests
	2.4 Detects	3.4 Establishes
	2.5 Compares	3.5 Reasons
	2.6 Classifies	3.6 Generalises
	2.7 Interpret	3.7 Predicts
	2.8 Explains	3.8 Judges

9. Affective Domain Society

(i) *Krathwhol et al.* (1964)

This taxonomy was developed by Krathwhol, Bloom and Masia (1964) in which internalisation was used as the organisational principle for the following levels (14).

1. Receiving	1.1 Awareness
	1.2 Willingness to receive
	1.3 Control or selected attention
2. Responding	2.1 Acquiescence
	2.2 Willingness
	2.3 Satisfaction
3. Valuing	3.1 Acceptance of value
	3.2 Preference for a value
	3.3 Commitment
4. Organisation	4.1 Conceptualisation of a value
	4.2 Organisation of a value
5. Characterisation	5.1 Generalised set
	5.2 Characterisation

(ii) *Hanna and Michaelis* (1977)

According to the authors, observing and recalling are basic processes involved for 'data gathering' which forms the foundation of the cognitive, psychomotor and affective processes. In

the affective domain, the 'attitudes and values' are categorised in the following manner in order of complexity of committment involved:

5. Integrating
↑
4. Preferring
↑
3. Accepting
↑
2. Complying
↑
1. Responding
↑

Data gathering
Observing and Recalling

This includes levels which describe the sequence of development of such interests, appreciations and adjustments as well as attitudes and values which schools try to encourage and develop in their students. Based on the principle of commitment, the levels are arranged in a hierarchical order.

10. Psychomotor Domain Taxonomies

(i) *Simpson's taxonomy* (1966-67) gives following levels for objectives arranged by the level of complexity and sequence in the performance of a motor act (15):

1.0 Perception	1.1 Sensory stimulation 1.2 Cue selection 1.3 Translation
↓	
2.0 Set	2.1 Mental set 2.2 Physical set 2.3 Emotional set
↓	
3.0 Guided response	3.1 Imitation 3.2 Trial and error
↓	
4.0 Mechanism	4.1 Habitual response 4.2 Patterned response
↓	
5.0 Complex overt response	5.1 Resolution of uncertainty 5.2 Automatic performance
↓	
6.0 Adapting and originating	(Suggested as a possible sixth level)

(ii) *Alles* (1967)

A major ordering principle in this taxonomy is 'routinisation' and adaptability is the second complementary criterion. Thus the principle of adaptive routinisation is significant for learning of psychomotor skills. Accordingly, the following set of ordered levels of execution in the psychomotor domain emerge (16).

1.00 Initiatory level of execution
2.00 Pre-routine level of execution
  2.10 Pre-routine Non-adaptive sub-level
  2.20 Pre-routine Adaptive sub-level
3.00 Routinised level of execution
  3.10 Routinised Non-adaptive sub-level
  3.20 Routinised Adaptive sub-level

(iii) *Dave* (1971)

He included the following levels arranged in terms of the concept of coordination (17):

1.0 Imitation	1.1 Impulsion 1.2 Over repetition
↓	
2.0 Manipulation	2.1 Following precautions 2.2 Selection 2.3 Fixation
↓	
3.0 Precision	3.1 Production 3.2 Control
↓	
4.0 Articulation	4.1 Sequence 4.2 Harmony
↓	
5.0 Naturalisation	5.1 Automatism 5.2 Internationalisation

(iv) *Harrow* (1972)

He developed a taxonomy to categorise movement behaviours according to the following levels (18):

1.0 Reflex movements
  1.1 Segmental
  1.2 Intersegmental
  1.3 Supra segmental

2.0 Basic fundamental movements	2.1 Locomotor
	2.2 Non-locomotor
	2.3 Manipulative
3.0 Perceptual abilities	3.1 Kinesthetic
	3.2 Visual
	3.3 Auditory
	3.4 Tactile
	3.5 Coordinated
4.0 Physical abilities	4.1 Endurance
	4.2 Strength
	4.3 Flexibility
	4.4 Agility
5.0 Skilled movements	5.1 Simple
	5.2 Compound
	5.3 Complex adaptive skill
6.0 Non-discursive communication	6.1 Expressive movements
	6.2 Interpretive movements

(v) *Hanna and Michaelis* (1977)

Like cognitive and affective domains, the levels categorised in skill emerge from the foundational category of data gathering through observing and recalling. Increasing independence from assistance or direction in the performance of an act constitutes the hierarchy of instruction as shown below (19):

5. Improvising
↑
4. Applying
↑
3. Mastering
↑
2. Patterning
↑
1. Imitating
↑

Data gathering
Observing and Recalling

(vi) *Maclay's taxonomy* (1969)

This taxonomy is based on the work of E.J. Simpson (1966-67) and comprises the following levels (20).

1.0 *Perception*
- 1.1 Sensory stimulation
- 1.2 Static cue discrimination
- 1.3 Dynamic cue discrimination
- 1.4 Translation

2.0 *Readiness* for further experience following perception (1.0)
- 2.1 Mental readiness to select and synthesise
- 2.2 Physical readiness to adapt
- 2.3 Emotional readiness to respond

3.0 *Guided response*
- 3.1 Imitation
- 3.2 Trial and error
- 3.3 Following of written or other guidance
- 3.4 Obtaining a series of simple experimental results

4.0 *Mechanism*
- 4.1 Choice of materials, apparatus, tools, reagent etc.
- 4.2 Planning the order of activities
- 4.3 Exercising skill
- 4.4 Execution of task
- 4.5 Interpolation and/or prediction from results

5.0 *Complex response*
- 5.1 Adaptation
- 5.2 Competent planning and confident procedure
- 5.3 Skilled execution and professional poise
- 5.4 Clear and concise reporting
- 5.5 Prediction of possible extended investigations

6.0 *Complex response*
As under 5.0 but without guidance

11. Perception Domain

Moore (1967)

In addition to cognitive, affective and psychomotor domains, Moore (1967) identified the need for a fourth domain, the perceptual. According to this domain, the following levels arranged

in terms of integration are given (21):

1.0 Sensation	1.1 Visual	Awareness of qualities of stimulus
	1.2 Auditory	
	1.3 Tactile etc.	
2.0 Figure perception	2.1 Size	Figure-ground
	2.2 Form	
	2.3 Location	
	2.4 Position	
	etc.	
3.0 Symbol perception	3.1 Letters	
	3.2 Digits	
	3.3 Other signs	
4.0 Perception of meaning	4.1 Value of percept or symbol	
	4.2 Relations such as	
	(a) cause-effect	
	(b) symbol-percept	
	4.3 Abilities to generalise	
	4.4 Understand implications	
	4.5 Take decisions	
5.0 Perceptive performance	5.1 Observation	
	5.2 Diagnostic ability	
	(a) Medical problems	
	(b) Artistic products etc.	
	(c) Electrical systems	
	5.3 Problem solving	
	5.4 Creativity	

12. Experiential Domain

(i) *Steinaker and Bell* (1975)

They have developed a model for class room experience though the experiential taxonomy. This is based on Gestalt's approach and is called a functional tool by the authors of the taxonomy.

All the three domains are integrated in this taxonomy. Its major aspects are:

1.0 Identification of components	1.1 Motivation
	1.2 Interaction
	1.3 Assimilation
	1.4 Replication
	1.5 Involvement

2.0 Development of curricular experiences	2.1 Terminal performance objectives
	2.2 Introduction and reinforcement
	2.3 Mastery sequence
	2.4 Strategies
	2.5 Skills
	2.6 Evaluation
3.0 Format of Evaluation	

All the taxonomies discussed above are useful in

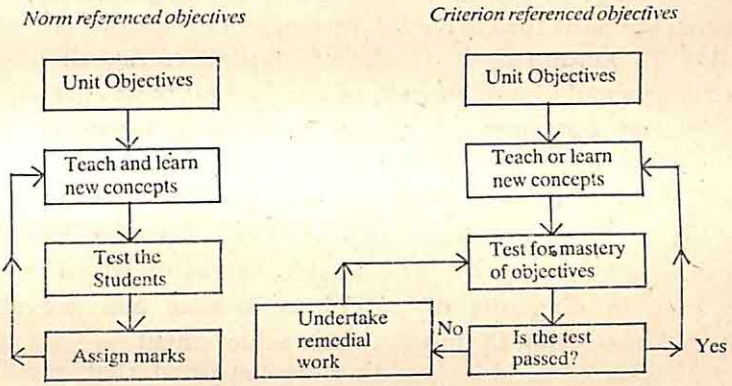
(a) preparing objectives, planning instruction and assessment.
(b) identifying abilities implied under each objective.
(c) developing curriculum content and methodologies of instruction.
(d) item banking.
(e) understanding relationships among objectives of different categories in various domains.
(f) reorientation of courses of studies in teacher education.
(g) developing objective-based evaluation procedures.

13. Criterion Versus Norm-referenced Objectives

We have already defined objective as an intended product of learning. This is reflected in the performance of students. While writing objectives we may specify a *criterion* of acceptable performance. According to Mager (1962) we should not only say what the learner should be able to do but also how well he should be able to do it before his performance is acceptable. For example, "given twenty three digit multiplication sums the students should be able to solve at least 15 correctly within half an hour", not only describe the students' behaviour but also the level of performance to "pass". It is presumed that all those students who score less than 15 marks need remediation before they proceed to the next unit of learning.

This idea is comparable to the concept of a *mastery test*. A mastery test covers vital essentials of a course which are represented in its items and are considered essential for students to know before they proceed to the next level of instruction. The student must demonstrate that he has "mastered" certain basic objectives at a predetermined level of performance. Such mastery

tests distinguish between students who have mastered and those who have not mastered the basic essentials. In contrast to this, *discrimination tests* which we normally use in our class rooms are designed to differentiate students at different levels of achievement irrespective of the standard of achievement. Thus the criterion-referenced approach is like the programmed instruction approach in which a student masters a small unit of materials. He then proceeds to the next one or if need be, does remedial work to master it and then proceeds to the next unit. The two approaches can be appreciated on the following models, one representing the traditional norm-referenced and the other criterion-referenced approach.



In the criterion-referenced model it is, therefore, imperative to describe objectives in terms of criteria regarded as an acceptable standard of performance which must be reached before the student is allowed to proceed to the next learning unit. A teacher, therefore, must appreciate how stating of objective in behaviour terms, non-behaviour terms and criterion-referenced terms is useful or is a handicap in evaluating students and for feedback of evidence for instructional practices. Three examples of stating objectives will further clarify this.

1. *Non-behavioural*

To use correct English grammar.

2. *Behavioural*
Given sentences containing grammatical errors, the students will be able to rewrite correctly.

3. *Criterion-referenced*
Given sentences containing grammatical errors, the student will be able to rewrite correctly at least 80 per cent of the sentences.

14. Place of Objectives in Teaching and Testing

Since specification of objectives is primarily meant for visualising the scope and nature of each objective, it helps the teachers and evaluators to focus their attention on the terminal behaviours of students. It is these intended learning outcomes that become the basis for evaluation as well as instruction. As far as testing is concerned, it is almost imperative that the testing situations should be so selected as to cover all these behaviours or learning outcomes or at least an adequate sample of those behaviours. Each of these behaviours can be tested singly or in combination of two's or three's as we normally do by using essay-type questions. Such a possibility of testing learning-outcomes one by one is quite desirable, especially when we are interested in diagnosis of pupils' weaknesses and providing remedial instruction to improve their achievement.

On the other hand it should not be construed that teaching can also be done on one-to-one basis. For example, it is quite ridiculous to say, "I am now teaching for interpretation, then for analysis and then for developing ability to hypothesise." If we purposely do this, it means that we teach for a specific behaviour and test for that behaviour. We teach for another behaviour followed by testing for the same behaviour. Obviously, if we do that we are teaching at a low level. In reality this is not the case because teaching is an integrated act. It is neither possible, nor desirable to visualise that a teacher is teaching for one particular objective for ten minutes, for the second objective for five minutes and so on. This is not the intent of specifying objectives in behavioural terms. What is needed is to appreciate the relationship between objective, teaching or learning and testing. An examination is a sampling

process while teaching is an integrative process. But what is worth realisation is the objective-based teaching and objective-based testing so that the one could be geared to the realisation of objectives and the other to the testing for those objectives. How both teaching and testing can be dove-tailed to the point of instructional objectives viz-a-viz unit objectives, is made clear in the unit plan 14.1 given on page 138-39.

14.2 *Objective-based Questions Testing Specific Objectives*

Concept

For germination of seeds, air, water and suitable temperature are necessary.

Specific objectives that can be tested.

1.0 *Knowledge*

1.1 *Recalls*

List the three conditions necessary for germination of seeds.

1.2 *Recognises*

Seed can germinate in the absence of:

A. Moisture, B. Temperature, C. Oxygen, D. Soil.

2.0 *Understanding*

2.1 *Translates*

Define the term 'germination' in your own words.

2.2 *Interprets*

Three beakers as shown below were placed at room temperature for a week. What does this experiment demonstrate?

Figure-A

Figure-B

Figure-C

Teaching point: Conditions necessary for germination

Objectives	Teaching-learning activities	Evaluation questions
1. The pupil recalls the conditions necessary for germination.	1.1 The teacher explains the necessity of air, water and suitable temperature for germination. 1.2 The pupil takes notes and reads the text.	1.1 Describe an experiment to show that temperature, air and water are necessary for germination.
2. The pupil develops skill in setting up the three-bean experiment.	2.1 Teacher gives instructions and precautions for setting up the experiment. 2.2 The pupil fits up the apparatus and performs the experiment.	2.1 Fit up the apparatus to demonstrate the three-bean experiment.
3. The pupil draws the diagram of the experimental set up.	3.1 The pupil draws the sketch of the experimental set up and labels the various parts.	3.1 Draw a labelled sketch to illustrate the three-bean experiment.
4. Pupil compares the conditions available to the three seeds used in the experiment.	4.1 Teacher puts questions like the following: (a) Why do we use three seeds in the experiment? (b) Why do we keep the middle seed half immersed in water? 4.2 The pupils try to answer the questions posed.	4.1 Why are air and moisture necessary for germination of seeds? 4.2 What will be the inference if we do not use the seed which is kept under-water?
5. The pupil develops the ability to apply his knowledge of	5.1 The teacher poses questions like the following:	5.1 Why is dry grass spread on nursery beds when it is very

conditions for germination in explaining phenomena involving a unfamiliar situation.	(a) Why do maize seeds not germinate when sown deep in the soil? (b) If there is heavy rain immediately after sowing of the seeds they do not germinate. Why?	cold?
	5.2 The pupil analyses the given situation and gives various responses to the question posed.	5.2 A gardner puts sweet-pea seeds in the flower bed under optimum conditions of temperature but they did not germinate. What could be the possible reasons for this?
	5.3 The pupil asks questions for further clarification.	

Note: This plan is only suggestive and not exhaustive. The main purpose is to appreciate the integral relationship among the three major components of the teaching-learning process.

2.3 *Illustrates*

Give one example of a seed that requires more moisture for germination and one requiring comparatively less moisture.

2.4 *Identifies relationship*

Why does a seed not germinate when sown too deep?

2.5 *Compares*

In what respect does germination of a gram seed differ from that of a castor seed?

2.6 *Locates error*

Observe the given diagram C of the experiment (Item 2.2) to demonstrate the conditions necessary for germination and point out the mistake if any. Suggest rectification also.

2.7 *Explains*

Explain in not more than 100 words the mode of germination in a pea seed.

2.8 *Classifies*

Categorise the following seeds into two groups on the basis of their mode of germination.

A. Pea. B. Gram. C. Castor. D. Maize. E. Sun Flower. F. Onion.

3.0 *Application*

3.1 *Analysis*

Some quality seeds of wheat were sown in two separate pots A and B using the same quality of seeds and sample of moist soil and kept at the same place. The seeds in pot A germinated whereas the seeds did not germinate in pot B. Under what conditions would this experience be relevant?

3.2 *Hypothesis*

When mustard seeds are scattered on a suitably prepared soil, they germinate but maize seeds do not. But when

sown a little deep under the same soil, maize seeds germinate but not the mustard seeds.

3.3 *Suggests procedure*
Bigger seeds require more moisture than smaller seeds. Suggest an experimental procedure to test this hypothesis.

3.4 *Gives reason*
Why seeds generally do not germinate when there is a heavy rainfall immediately after sowing?

3.5 *Draws conclusion*
What conclusion can you draw from the observations of experiment shown in figure C, item 2.2.

3.6 *Predicts*
What will happen to germination of seeds in figure B, (item 2.2) if the whole apparatus is removed from a lighted room to a dark room?

4.0 *Skill*

4.1 *Draws sketches*
Draw a labelled diagram to show the experimental set-up of the three-bean experiment.

These examples indicate how the same concepts can be used to test different specific objectives.

15. Objectives and the Item Bank

Development of an item bank cannot be thought of without the knowledge of instructional objectives. In fact each and every question in a bank is objective-based i.e. testing one or the other objective. The only thing that a teacher should be cautious of is the cognizance of the objective or the behavioural outcomes which each question attempt to test. In fact the quality of question bank determined by the objectives tested by questions which form the bank. Collection of questions testing the merely knowledge of facts do not make a quality bank, whereas

a bank with a majority of its questions testing students' ability to apply, analyse or evaluate is definitely the superior question bank. The greater the variety of question under each objective i.e. the more the coverage of specified behaviours implied by an objective, the more the validity of such a question bank would be for the purpose of testing. It is, therefore, necessary that teachers try to frame questions on selected concepts testing different specific objectives (specifications) implied under various objectives as exemplified above under 14.2.

16. Objectives as Help or Hindrance

There is an overwhelming belief in the need for clarity and specificity in stating educational objectives. The three major reasons are: they provide the direction or goal towards which a curriculum is aimed; facilitate selection and organisation of content and learning experiences and become a referent for evaluation of the learning outcomes. As early as 1924, Bobbit (21) in his book, *How to make a Curriculum*, listed nine areas in which 160 major educational objectives were mentioned. Pendalton listed 1581 social objectives for English; Giller listed more than 300 for Arithmetic in grade 1-6 and Billings prescribed 888 generalisations for social studies. By 1930, however, this movement collapsed due to the non-manageability in classroom of such long lists by the teachers. Again in the late forties and during the fifties curriculum specialists started emphasising the need for specific educational objectives. Herrick (1950), Tyler (1951), Bloom (1956), Krathwhol (1964), Harrow (1972) were the outstanding curriculum specialists who gave their classification or taxonomies.

Inspite of the tremendous work done on this direction specialists like Eisner (1967) are very critical of such specified objectives formulated in advance. He argues that the process of instruction is very complex and dynamic rather than mechanistic with numerous outcomes which cannot be listed in behavioural and content terms in advance (22). One cannot predict in advance many of the collateral or emergent outcomes. Evaluation experts of today do not however claim that a priori guesses can cover the entire outcomes of learning. All learning

outcomes which are amenable to description and identification can be formulated and agreed upon. A statement of objectives in operational terms does provide the needed framework and basis for the teaching-learning process. What is perhaps more agreeable is to start with the rational formulation of objectives which may be called 'hypothesised objectives' or construction objectives. At the end of instruction, some of them might be modified, deleted or added, thereby leading to developed or summative objectives. This middle-of-the-road approach of hypothesised and emergent objectives can better claim to explain and reconcile the two view points.

A second hindrance is the constraints that a discipline puts. It is alleged that more precision in specification of objectives is possible in science subjects than in arts, where it may be difficult or not even desirable to accept the same end-product. Novelty, creativity and originality are unpredictable. This is also challenged by Atkin (1963) according to whom (23) the so-called relative simplicity in science and mathematics is illusory if we examine the latest curriculum projects in Science and Mathematics. Nevertheless, to my mind the nature and structure of a discipline does create problems regarding agreement of the intended outcomes of learning in case of less matured disciplines like History and Sociology as compared to Mathematics and Physics.

The third hindrance that Eisner points to is that objectives cannot be used as criteria or standards against which achievement is measured. Making a distinction between the application of a standard and the making of judgement, he says that socially defined standards as in the case of logic, rules of grammar it is possible to measure and compare quantitative and not qualitative judgements that can be used to measure aesthetic objects or reflective essays and the standards are not applicable. This comment, as Hastings points out is based on a wrong concept of evaluation which is equated with testing. All description about pupils' performance are not quantitative but are qualitative also. So evaluation consists of descriptions and judgement of those descriptions. That descriptions should only be quantitative is indeed harmful.

Another hindrance pointed out is that formulation of objec-

tives prior to curriculum activities is not psychologically the most efficient way to proceed. Teachers can identify first the appropriate activities, rich in educational potentials, and then identify the related objectives. Macdonald also supports this view by saying that this is a heuristic device of formulating objectives which are known only after completion of an act and not before: "What am I going to do or teach?" And not with, "What am I trying to accomplish?" Addition, deletion and modification of objectives is the strategy in all modern projects on curriculum in U.K. Stated objectives, during instructional process get metamorphosed into new outcomes. But this is possible only when some advance list of objectives is formulated on rational basis and experience. Such constructed objectives may later on be transferred into formative objectives during the teaching learning process and finally come out as summative or emergent objectives.

Thus this controversy of specific objectives is more apparent than real. If one accepts evaluation as a comprehensive concept, involving judgements on both the qualitative and quantitative aspects and develops faith in the goal-directedness of educational efforts, formulation and specification of educational objectives would not only be useful in planning and developing curriculum content and methodology but also in validating the whole process of teaching and learning by reflecting the intended learning outcome. The only caution is that too much fragmentation into unwieldly lists should be avoided to make it more realistic and practicable in the classroom. Moreover such a list should be considered only tentative and must be revalidated in the light of instructional outcomes resulting from the pupil-teacher interaction. Once objectives are clearly formulated, the evaluation process becomes more well directed.

CHAPTER V

MODERN EVALUATION—A TOTAL SCHOOL CONCEPT

1. Introduction

The use of words like measurement, examination and tests have been in vogue ever since the formalisation of education to judge the effect of instruction on students' achievement. The subjective judgement of the teacher, classroom questioning, performance on a school assignment, quality of home work, pupils' products in the form of a chart, model, or a composition and observation of behaviour of students are the modes of assessment practised by teachers formally or informally twoards instructional ends. The focus of all these methods of pupil's assessment is always on the improvement of student's learning. Since written work occupies a far more prominent place than oral and practical work, the use of written examinations in the form of periodical, terminal or annual tests have come to stay in our educational system. More often than not, such written tests are used for judging the end-product of learning rather than as a teaching-learning device to improve student's achievement. The reason is obvious. The evaluation concept is considered synonymous with an examination or a test. This has narrowed down the scope of evaluation which in fact is a much broader concept and is integral to the total teaching-learning act. Thus the nature of the evaluative process can better be appreciated if evaluation is considered a broad, dynamic and a composite concept.

It is, therefore, intended that study of this chapter should enable the reader to

(a) define the various terms like examination, measurement assessment, evaluation etc. in his own words.
(b) differentiate the nature, scope and role of diagnostic formative and summative evaluation in teaching-learning process.
(c) identify relationship of evaluation with other components of the teaching-learning process.
(d) interpret the basic principles of evaluation and their implications.
(e) explain the basic characteristics of evaluation in the context of teaching-learning process.
(f) consider evaluation as a total school concept rather than occasional act of measurement of pupils' attainments only.

2. Nature and Scope of Evaluation

Various terms associated with evaluation have already been described in the first chapter and need not be repeated. An attempt can be made now to visualise the relationship among the more common and interchangeably used terms on the basis of the scope and function of each. Suppose we are interested in judging the effectiveness of our instructional efforts in terms of the impact on students' learning or achievements we might have to do the following:

(a) The type of learning in the measurement of which we are interested—intellectual, affective or psychomotor development.
(b) Cognizance of acceptable standards of performance regarded as intended learning.
(c) The hypothesis to be tested in the process to serve as a proof of evidence having a bearing on the expected learning.
(d) Process of collecting evidence about the expected learning (examination, observation, opinion, products, analysis, ect.)
(e) The use of instruments appropriate to the evidence to be collected.
(f) The use of criteria for appraisal of evidence (estimation or accurate measurement).
(g) Making a value-judgement in terms of given criteria.

Expected end-product of learning, performance standard, indicators of acceptable standards, mode of collecting evidences, data gathering instruments, appraisal of evidence and judgement making are, therefore, the corresponding nodal points in evaluation. This leads us now to defining the related terms to determine the nature and scope of each.

(a) *Achievement* refers to the intended learning in cognitive, affective or psychomotor domain regarded as the end-products of learning, reflecting the desired standard of performance (1).

(b) *Test* refers to a product of a device used for determining the truth or falsity, meaning or value of a hypothesis (2). A Question Paper, a unit test or a term test is a tool or an instrument of measurement which is used to collect the evidence related to intellectual, affective or psychomotor development (learning).

(c) *Examination* refers to the process of collecting evidences about pupils' achievement. A process may be a written, oral or practical examination, an observational or inquiry technique, testing or analysis of a product or a document etc. Thus testing is one technique of examining like the observational technique (3).

(d) *Assessment* refers to teacher's estimate of a value against a certain standard. Therefore, it is only an estimate and not an accurate evaluation. The term is appropriate in relation to such traits as attitudes, personality, co-curricular activities, S.U.P.W. where qualitative measurement (or non-measurement) is called for, unlike written testing which lends itself better to quantitative measurement (4).

(e) *Appraisal* refers to accurate valuation of some characteristic or attributes against some criteria which are usually developed by experts or knowledgeable persons in the field (5).

(f) *Evaluation* refers to the process of collecting information about a pupil's achievement through measurement (quantitative) or non-measurement (qualitative) on the basis of which judgements are formed and decisions made. Thus information gathering, forming judgements and making decisions are the major components of evaluation (6).

(g) *Measurement* refers to the act of ascertaining the extent

of pupil's growth or achievement. It deals with comparison of a quantity with an appropriate scale for the purpose of determining the numerical value on a scale that corresponds to the quantity to be measured. Thus it is an act of assigning numerals (7) to an object according to certain rules (Lindquist). Relationship between measurement and evaluation can be visualised by the following equation (by Gronlund, 1965).

Evaluation = measurement + value judgement

Thus there cannot be any evaluation without measurement but both are not synonymous. The following examples would illustrate the relationship among various terms mentioned above.

(i) A unit *test* was used to measure students' learning on the unit 'Respiration'. (Tool)
(ii) The *observation* of the teacher was the basis of judging students' behaviour in the morning assembly. (Technique)
(iii) On the basis of *testing*, students were identified for remedial teaching. (Technique)
(iv) The annual *examination* alone was made the basis of promotion of students. (Process)
(v) Most of the students were *estimated* as having good health habits during the last term. (Assessment)
(vi) Alka secured *58* marks out of 100 in English. (Measurement-quantitative)
(vii) Kumar got *excellent* in Hindi dictation. (Measurement-qualitative)
(viii) In terms of *mean performance* (45) of the class Alka is ranked among the top five students of her class. (Appraisal)
(ix) Luna is the *brightest* boy in English and deserves a prize. (Evaluation-judgement forming)
(x) Sanjay is one of the *weakest* students in the biology class and should be *given remedial instruction.* (Evaluation—decision-making)

The term evaluation is, therefore, no longer used as equivalent to examination, measurement or assessment. Evaluation is a process which is concerned ultimately with decision-making.

Such decisions may be about the process of learning, testing, administration of instruction or the end-product of learning. For decision-making, formation of sound judgements is a pre-requisite. Without valid judgements proper decision cannot be taken. Judgements are formed in terms of certain referents, purpose and methodology of comparison of performance. Judgements in turn are based on the information collected about students. Collection of information is, therefore, the basic ingredient of evaluation.

Information gathering, information processing, judgement forming and decision-making are, therefore the four major components of evaluation. We may now define evaluation as under:

> Evaluation is a process of collecting information about pupils' learning and processing it to form judgements which in turn are used to take decisions.

3. Diagnostic, Formative and Summative Evaluation

Whenever a prefix is added to evaluation, it restricts its scope. As such, diagnostic, formative and summative evaluation refer to only one phase of evaluation, emphasising the diagnostic, developmental and judgemental aspects of evaluation.

Diagnostic evaluation is primarily meant for placement of students on the basis of pre-requisite entry behaviours. It determines the cause of repeated difficulties of pupils so as to provide for remedial measures. It is used before the actual teaching of the unit in hand as well as during developmental teaching when pupils fail to grasp the subject. The focus of diagnostic evaluation is on the identification of trouble spots and their causes so that a deliberate attempt can be made to improve pupil's ability to benefit maximally from instruction. Therefore, if the purpose of evaluation is to place the individual on a learning continuum so as to enable him to benefit from instruction, it is diagnostic evaluation.

Formative evaluation is an integral part of the teaching-learning process and therefore, cannot be considered in isolation

or apart from instruction. The focus of formative evaluation is on improvement of students' learning and not on certification of achievement. It refers to continuous evaluation by means of unit tests, informal class tests, assignments, classroom questions etc. It helps to diagnose pupils' strengths and weaknesses. The results of formative evaluation are used for further improvement of instruction and pupils' achievement by adjusting ends and means. Classroom instruction ordinarily invokes formative evaluation. Thus, formative evaluation is diagnostic on the one hand and developmental on the other.

Summative evaluation is considered as an end-of-the-term or end-of-the-course activity. It is directed towards making judgement about the end products of learning and is usually undertaken in a formal manner. The focus of summative evaluation is on the measurement of pupils' achievement and its certification. Term tests, annual tests and external examinations, whether conducted by schools or a public agency, can be categorised as summative tests. The purpose of summative evaluation is classification and grading of students rather than identification of pupils' strengths and weaknesses. The results of summative evaluation are used for certification and/or passing judgement on students' achievement.

The three types of evaluation are more or less inter-dependent, representing the three major phases of evaluation in the teaching-learning process. The results of summative evaluation can be used after analysis, for a diagnostic purpose and for taking administrative decisions in re-classification of students or for instructional changes. The results of formative evaluation can similarly be used for classifying students in to different groups for remedial work. Likewise, the results of diagnostic evaluation, whether before-instruction or after-instruction can be used not only for placement purposes but also for grading students into mastery, partial mastery, and non-mastery groups for the purpose of remedial programmes. Therefore, it is mainly the purpose and the use of the results of evaluation that determines the premium to be laid on one or the other ttype of evaluation. The following table would further clarify he relationship among the three types of evaluation (8):

Basis	Summative	Formative	Diagnostic
1. Nature	Judgemental	Developmental	Descriptive
2. Focus	Measurement of achievement	Improvement of achievement	Discovering of deficiencies
3. Major functions	Certification and grading	Location of errors—Feedback—to learner & teacher	Proper placement for instruction
4. Timing	End of unit, term, course	During teaching learning process	Before actual instruction and after analysis of test results
5. Aspect of evaluation	Cognitive, psychomotor	Cognitive	Cognitive, Affective, Psychomotor. Phys. Psy and environmental factors
6. Instrumentation	Written exam, oral, practical	Specially designed tests	—Formative and summative tests —Stand-Ach tests —Teacher-made instruments —Obs—Checklists
7. Sampling of objective/content	Related to weighted objectives/course units	Specific to relevant tasks in hierarchy	—Specific to educational behaviour —Specific to non-educational factors
8. Analysis (Item difficulty)	Average 35-70%	Cannot be prescribed	Essay 65% or higher
9. Interpretation	Generally in terms of a group	In terms of criterion	In terms of both norm and criterion
10. Reporting	Total scores or sub-scores by objectives	Pattern of pass-fail scores on each hierarchical task	Individual profiles by sub-skills

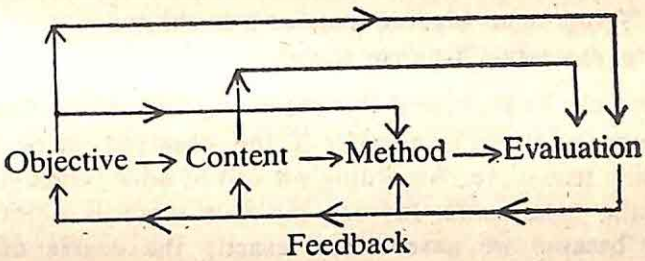
4. Evaluation—An Integral Part of the School Curriculum

More often than not, evaluation is considered as an end-of-the-course activity rather than a part and parcel of the total curriculum. One reason for this is the narrow concept of curriculum conceived by teachers. Without going into the nuts and bolts of the curriculum, it may suffice here to say that "curriculum is the totality of learning experiences for which a school undertakes the responsibility to reach predetermined goals through continuous appraisal and adaptation of the end products of learning and the process of learning". From this definition we may identify five fundamental questions which a teacher can ask himself. These are:

(a) Why should I teach a particular subject (or a unit)?
(b) What should I teach in that subject (or unit)?
(c) How should I teach it well? (or how best can students learn?)
(d) How best have I taught it? (or how best students have learnt?)
(e) In what ways, I can improve my teaching? (or students' learning?)

The first question refers to instructional objectives which involve the question of values, preferences desirability and purposefulness and, therefore, reflects the philosophical basis of the teaching-learning process. The second question relates to knowledge in the discipline that forms the warp and woof of the subject matter i.e. the content of teaching or learning. Content selection involves purposes, social needs and priorities which the society demands. As such, this aspect rests on a sociological basis. The third question relates to teaching-learning strategies the selection of which depends on both objectives and the content but are conditioned by the needs of the learner and the psychology of learning and instruction. Thus selection and organisation of learning experiences are governed mostly on a psychological basis. The fourth question refers to the performance assessment which depends on the first three aspects; namely, the objectives, content and methodology

of teaching. Validity and reliability are basic attributes of any instrument of evaluation for the collection of data, on the basis of which judgements are formed. Thus the evaluation aspect involves a scientific basis of measument. The last question relates to improvement. This requires regular feedback of evidence that accrues from the evaluation process. This involves diagnosis and remediation and adjusting of ends and means to improve students' learning. Thus feedback brings in dynamism by generating interaction among all the components of the curriculum and may, therefore, be termed as the ecological basis of the teaching-learning process. The relationship among the five aspects can be depicted in the form of following diagram:



5. Basic Principles of Evaluation

There are certain assumptions about the individual learner and the evidence that we get about him. It is on these assumptions that principles of evaluation can be identified. These assumptions are as under:

(a) *Learner has measureable ability*

According to the measurement theory an individual has some real ability, may be intelligence, or the ability to apply knowledge in a new situation, that differs from individual to individual. Thus each learner has some amount of this ability which can be measured. A learner's true score or true ability is estimated from the observations or evidences collected through measurement. Thus a learner has a true score.

(b) *An observation contains error*

A score obtained from an observation is not the same as the true score because whenever we measure we make errors. If the height of a person is measured 3 or 4 times with the same tape by different persons there would be some difference, may be 1/24 or 1/36 of a centimeter. When this error is possible in case of physical measurement, it is more so in case of human or psychological measurement.

(c) *A true score is an observed score minus error score*

In terms of the measurement theory this assumption is written algebraically as under:

$$X=t+e$$

where X represents the observed score of the learner,
't' represents his true score or true ability,
'e' represents his error score.

Error can be positive if the observed score is more than the true score and it can be negative if the observed score is less than the true score. Accordingly it will be added or subtracted from the true score. Indeed, the true score is never truly known because we never know exactly the degree of error involved in a given case. Suppose we do know that Raman's IQ is 110 but when we measure it may be 115 or 105 i.e. the score he gets may range from 105 to 115. In these two cases we may write the score as under, using the same formula:

$X=t+e$ $(115=110+5)$ OR $X=t+e$ $(105=110+(-5))$.

Thus identification of errors, their prevention, estimation and then tolerating these errors, is basic to the principles of measurement. Such errors, called measurement errors, decrease the accuracy of the observed scores. The error which are associated with traditional measurement errors (unreliable scores) refer to unreliability while those which occur due to information that is inappropriate for judgements refer to invalidity.

Some basic principles help us to cope up with the problem of errors. TenBrink, (1975) has listed four basic principles (9).

First Principle

5.1. *Errors can be Identified*

Many sources of errors can be identified and an effort made to prevent their recurrence. The following sources contribute to such errors.

(a) *Errors in information gathering instruments*
These may be due to:

(i) inappropriate content,
(ii) inappropriate sample of content,
(iii) too difficult or too easy items,
(iv) non-comprehensibility of the items,
(v) ambiguity in the directions of items,
(vi) lack of information about the length of test, time spent on observing etc.,
(vii) overemphasis or underemphasis on objectives.

(b) *Errors within the information gathering process*
These may be due to:

(i) unclear directions in test administration.
(ii) physical conditions for administration.
(iii) technical incompetence of invigilators.
(iv) computational and other errors in scoring.
(v) tabulation and recording errors.

(c) *Errors within the evaluee himself*
These may be due to:

(i) fluctuality of human traits e.g. attitudes, interests etc.,
(ii) specific attitude towards the task being evaluated.
(iii) misunderstanding of directions and other task related to what is expected of him,
(iv) test wiseness or unwiseness of the test,
(v) particular response set i.e. tendency to respond in a consistent fashion like tendency to guess,

(vi) guessing,
(vii) test-taking ability (or inability),
(viii) state of health of the individual, physical, mental or emotional.

Second Principle

5.2. *Errors can be Minimised*

The more the errors presented the more valid and reliable is the information we are likely to obtain. This can be ensured by:

(i) relating tasks to purpose of evaluation e.g. construction, selection, scoring, judgement making, use of evidence etc. features which should all be based on the reason or purpose of testing,
(ii) avoiding ambiguity and inconsistency in the use of language,
(iii) giving clear and consise instructions/directions,
(iv) obtaining representative sample of information which describe the population, choosing appropriate technique of sampling and obtaining the sample.

Third Principle

5.3. *Errors can be Estimated*

If some sort of estimation is made about the amount of errors in the measurement data we can use it to temper judgements and improve decision-making. To determine the magnitude of errors we can use two strategies. One is to get empirical evidence and second is circumstantial evidence or searching for clues about the validity of responses. For this various measures can be adopted such as the following.

(a) *For reliability*
(i) Coefficient of stability or consistency over time.
(ii) Equivalence or consistency across different forms.
(iii) Equivalence and stability

(iv) Internal consistency or consistency across different parts of a test.
(v) Scorer (inter-judge) reliability or consistency across raters.

(b) *For validity*
(i) Concurrent validity.
(ii) Predictive validity.
(iii) Construct validity.

Fourth Principle

5.4. *Errors can be Tolerated* (lived with)

Some errors will always occur inspite of the best efforts of careful and a systematic way of obtaining information. Nevertheless, this information can still be used fruitfully if cautiously interpreted, checked and verified before use. For this we may 'hedge' or avoid committing oneself to substantiate the truth of the information.

(i) Take the observed score as an estimate of the true score which is accurate within certain limits.
(ii) Make a hypothesis on the basis of the information gathered so that the accuracy could be tested to confirm or reject in the light of additional information.
(iii) Confirm the accuracy or 'truth' of that score i.e. substantiate the information. Replicate the evidence to test the 'truth' of the informatinn (obtain information in similar circumstances) and supplement evidence for the testing truth of the evidence (obtain information under different circumstances).

6. Process Characteristics of Evaluation

A good evaluation programme has the following academic components or characteristics which reflect the evaluative process.

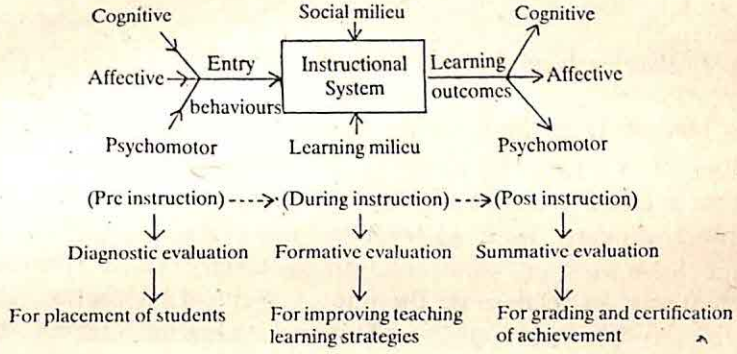
6.1. *Evaluation is Objective-centred*

Evaluation has a philosophical basis as it involves the question of values. What is worth learning or teaching is reflected in the formulation of objectives which may be intellectual, emotional, social, moral or physical in nature. Even when we accept a few or all of them, the question of preference and weightage comes in. Should we lay more stress on affective objectives than cognitive objectives at the primary stage? Should we teach moral values at this stage? Is socially useful productive work more important than the teaching of social studies? All such questions involve preference of values depending upon the idea of goodness or what is considered a desirable behaviour. In judgemental terms also, the evaluator has to be cognizant of the referrent he is going to use. Since judgements can be made in terms of a class, an individual or a performance standard, it is again a question of values, preference or purpose in the light of which judgements are to be made. As evaluation involves some value judgement, we have to make a choice among the competing values or cherished values which become the basis for obtaining information as well as for interpretation. It is, therefore, suggested that evaluation should be based on predetermined instructional objectives and the evidence validated against those very objectives.

6.2. *Evaluation is a Continuous Process*

In order to make evaluation a tool of learning it has to be used more frequently as a part of teaching-learning activities. Since an individual grows all the time and acquires new behaviours it becomes imperative that the desirability and adequacy of these behaviours be judged side by side. This is necessary to affect changes in teaching and learning and acquaint students with what is mastered by them and what is still to be mastered in order to reinforce learning and feedback the results for diagnosis and remediation. This continuity can be ensured by making performance-assessment a part of the teaching-learning activity. The more frequent the check-ups for regular feedback of students' performance, the more continuous does the evaluation process

become. Continuity of evaluation brings quality control into the process of learning. Therefore, evaluation should be made completely integrative with the teaching-learning process instead of being treated it as an end-of-the-course activity.



6.3. *Evaluation is a Comprehensive Process*

A pupil has different dimensions of growth; the intellectual, the emotional, the social and the physical. These aspects are represented in the form of cognitive, affective and psychomotor objectives. Therefore, unless evaluation provides evidence on all the aspects, viz-a-viz all the instructional objectives, it cannot be considered comprehensive enough to be depended upon. Likewise, coverage of both scholastic and non-scholastic areas is a must in any good programme of evaluation. Apart from testing all possible objectives and content areas, comprehensive evaluation involves the use of different tools and techniques to get different types of evidences on various aspects of evaluation. Evaluation, should therefore, be considered a comprehensive process.

6.4. *Evaluation is a Cooperative Process*

Since comprehensive evaluation seeks evidence on all aspects of pupils' development, the teacher alone cannot get full evidence about his growth. To collect evidence regarding social relationships, emotional behaviours, initiative, scientific attitudes,

social attitudes, likes and dislikes etc. We need the collaboration of pupils, his peers, his parents and all those teachers who watch him grow and develop. Therefore, for good evaluation the cooperation of different individuals and agencies is necessary and as such evaluation should be considered a cooperative process.

6.5. *Evaluation is an Integral Part of Instruction*

It is obvious from the teaching model referred to earlier, that evaluation is a part and parcel of the instructional process. It is not an end-of-the-course activity but is built into the teaching-learning process. Both instruction and evaluation go hand-in-hand. Such a type of evaluation is termed as a formative evaluation. Its focus is on the improvement of instruction viz.-a-viz. pupils' achievement. Unless evaluation serves this function through immediate feedback, there cannot be improvement in the teaching-learning process. Thus the role of evaluation as a teaching device should be appreciated, besides its judgemental role.

6.6. *Evaluation is a Dynamic Process*

Evaluation is based on objectives of instruction but at the same time it helps us to judge how far those objectives are attainable for a particular group of students. Likewise, evaluation is not only related to learning activities or instructional procedures but also provides evidence as to the effectiveness of the instructional efforts. Feedback loops from evaluation to the other components of the teaching model indicates how evaluation through regular feedback validates the whole teaching-learning process and leads to continuous improvement in a spiral fashion. As it is the feedback component of evaluation which has received little attention in our system of evaluation. Evaluation at present is concerned only with collecting data and its interpretation to a limited extent. But how to use the data for the improvement of learning or the learning process is not considered a useful actvity. But it is this aspect that helps us to judge the suitability of the objectives, the efficiency of the process of teaching or learning, the quality

of the measuring tool itself and even the accountability of the administrative decisions.

It is on the basis of feedback data that the learner can be informed of the adequacies to reinforce his learning. In the earlier stages extrinsic feedback in the form of encouragement, reward, incentive, etc. can be provided to the students. A regular feedback of all evaluation acts may ultimately lead to intrinsic feedback for the learner. Thus the most outstanding feature of evaluation is its ecological aspect which generates interaction among the various components of the teaching-learning process, thereby leading to the improvement in the learner, teacher, learning, teaching and even in administration. This interaction among various components of the educational process generated by evaluation through feedback mechanism, is indeed the spirit of modern evaluation. Hence, evaluation should be considered as a service component of the teaching-learning process to emphasize the use of evaluation results through regular feedback to improve teaching and learning.

7. Ecology of Educational Evaluation

From the earlier discussion about the nature, scope and characteristics of evaluation it is quite evident that the narrow view of evaluation which equated it with testing is no longer accepted. Evaluation is treated more as a process of decision-making than as a technique of testing. Its scope is, therefore, not limited to the evaluation of pupils' learning but is extended to the process of learning. As such, the focus of evaluation is more on the improvement of students' achievement than merely on measurement of their achievement.

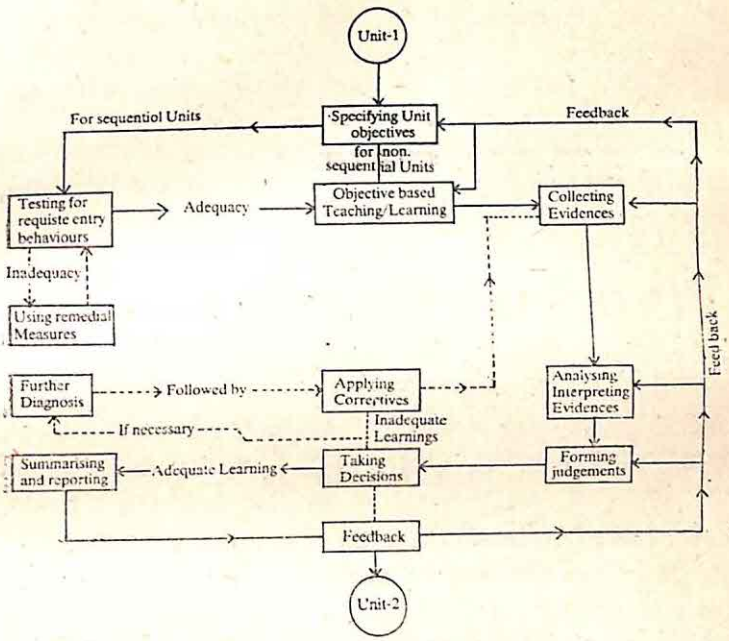
Since pupils' achievement is a function of the total school environment, it is necessary that the decision-making process must take into cognizance all conditions, antecedents and transactions which condition pupils' learning so that right judgements are formed. Thus the focus of modern evaluation is more on the description and interpretation of students' adequacies or inadequacies in learning than on the certification and prediction of their achievement.

Traditional evaluation emphasizes the judgemental role and

neglects the diagnostic role. Unless the evaluator is concerned with finding out the weaknesses in the pupils and the teaching-learning process itself, no administrative or academic decisions are possible to improve their achievement or learning. Evaluation, therefore, must be treated as an integral part of the teaching-learning process. In other words, evaluation should be treated as a intelligence service of the educational process. The focus of present evaluation is, therefore, not only on using evaluation as a device for the furtherence of pupils' learning but also for improving instruction.

Of late, there has been a tendency to link pupil evaluation with curriculum evaluation and programme evaluation. This is obviously due to the fact that the pupils' learning is conditioned by curriculum prescriptions and instructional efficiency and, therefore, pupils' evaluation is a part of curriculum evaluation which in turn becomes a part of programme evaluation that takes into account other variables like context, inputs, strategies, processes etc. (10). It is, therefore, necessary to appreciate this link and the interaction of the learner with the curriculum objectives, content, learning process, the teacher (instruction) and evaluation. It is for this reason that a focus of the modern evaluator on enabling the teacher to get more and more insight into the ecology of the evaluative process so that some sort of equilibrium could be ensured in the teaching-learning process. Thus if we say that evaluation is a total school concept, it would not be wrong.

The figure depicting integrative model of evaluation given on the next page would clearly indicate the place of evaluation in the teaching-learning process and its role in generating interaction among various components, thereby indicating the ecological basis of evaluation.



An Integrative Model of Evaluation

(vi) arranges ideas in a logical sequence.
(vii) uses proper gestures, avoiding mannerism etc.

(d) Develop criteria of assessment
For example:
(i) Vocabulary—10 marks
(ii) Pronunciation—10 marks
(iii) Grammatical correctness—10 marks
(iv) Fluency—5 marks
(v) Thought content—10 marks
(vi) Gesture and mannerism—5 marks

(e) Define each criterion
(i) Vocabulary (correct word, richness)
(ii) .....
(iii) .....
(iv) .....
(v) .....
(vi) .....

(f) Prepare a scale for each component
A five-point rating scale may be developed using 4-0 range of marks corresponding to better grades, A, B, C, D, E.
For example, for vocabulary we may define the limits as under:
(a) Always uses appropriate words.
(b) Frequently uses appropriate words.
(c) Usually uses appropriate words.
(d) Occasionally uses appropriate words.
(e) Seldom uses appropriate words.

(g) Develop marking scheme
On the basis of weightage given to various criteria and the rating of the examiner (A, B, C, D, E) on each criterion we can award marks. If weightage to vocabulary is 10 and the rating of a student on vocabulary is C, then $10 \times 2 = 20$, will be his score on vocabulary.

(h) Decide about the mode of conduct of examination
It is better to have a panel of 2-3 examiners. They may follow either holistic or analytical scoring. If the difference between the two approaches is not significant, then the botheration of the analytical method could be avoided.

Examiners may either choose one or more dimensions each and assess students with these dimensions only and then add the marks. Alternatively, each examiner may rate each student on each dimension and then pool the total marks for each student for comparison with the rating of other examiners.

It is advisable to have one internal examiner besides the two external examiners. This might help students to established rapport with the examiners and enable them to show their best.

(i) Devise questions and exercises

(i) Questions for establishing rapport (not to be used for assessment).

(ii) Questions of the short-answer variety like the guided conversation.

(iii) Questions requiring long responses. In this case the candidate is asked to select one out of 4-5 topics, gets 5-10 minutes for thinking and then speaking on the topic.

(j) Interpret the responses

This is important from the point of view of feedback to schools. An over-all global-type of summary information is not useful for feedback. Information on each dimension should be analysed and interpreted for feedback to the schools to enable them to appreciate the deficiencies and plan for remediation.

(k) Feedback the results for remediation

On the basis of detailed information provided by the examiners to the schools, relevant steps may be taken to provide a remedial programme to make up the deficiencies.

Work on oral testing in India appears to be merely on the theoretical side. A lot of empirical data are needed for judging the validity, reliability and practicability of oral tests especially for use in external examinations. With all its limitations, the place of oral examination in judging oral expression cannot be denied in the overall framework of evaluation, more particularly at the elementary stage.

CHAPTER VI

STEPS IN EVALUATIVE PROCESS

1. Introduction

The process of evaluation can be visualised in different ways depending upon the model of evaluation one seeks to propagate. If we recapitulate our discussion in previous chapter, we have tried to emphasise the integrative nature of evaluation. Evaluation is conceived as an integral part of the teaching-learning process rather than a judgemental act to be undertaken at the end of a unit, semester or a session. Likewise, our stress on its decision-making role must become our starting point and the feedback aspect as the last point in this process. Accordingly our sequential steps in explaining the evaluative process would be the specification of decision points, devising of situations, obtaining evidences, analysis and recording of evidences, forming judgements, summarising, reporting and feedback of information. It is, therefore, expected that by reading this chapter a teacher should be able to:

(a) specify the intended goals of learning for a unit,
(b) devise situations relevant to appraisal of intended outcomes of learning,
(c) collect the needed evidence using appropriate tools and techniques,
(d) analyse and record evidence in terms of purposes, goals and decisions,
(e) form accurate judgements and take appropriate decisions,

(f) summarise the evidence and report it meaningfully,
(g) feedback the evidence for reinforcement and applying correctives,
(h) appreciate the integrative model of evaluation.

Each of the above mentioned objectives can be achieved if the reader is able to understand the various tasks that are involved in the various steps discussed hereinafter.

2. Steps Involved

Step 1. Specifying intended outcomes of learning (decision points) In any programme of evaluation, the foremost decision that should be taken is the purpose for which evaluation is being undertaken. Is it selection, placement, treatment, certification of achievement, classification of students or to judge the effectiveness of the products, strategies of a programme? Accordingly, the object of decision may be persons, products or a programme. Will the evidence or data of evaluation be used for administrative, instructional, guidance or research purposes? When will the decisions be made; before, during or after instruction? Is the individual, the class or the institution that would receive priority in the decision-making process? Can the decisions so made be reversed, modified or are they irreversible? These are some of the significant decision points which are to be visualised before specifying the intended outcomes. In case of classroom evaluation, teachers are concerned mainly with the instructional objectives, and therefore, we may restrict our discussion to pupil evaluation in this chapter.

Since we have advocated the integral approach to teaching and testing, it is assumed that unit teaching and unit testing will form the basis of the evaluation approach. As such, specification of intended outcomes of the unit (topic) of teaching will be the first step in the process of evaluation. While formulating objectives for the unit of teaching, care may be taken to include all attainable objectives. Though affective and psychomotor objectives are important and complementary to cognitive objectives, yet for reasons of practical difficulty, from the point of view of their teaching

vis-a-vis evaluation, they are generally ignored, especially in external examinations. This is more true of affective objectives while practical examinations, whenever they are in vogue, do take care of psychomotor outcomes.

The instructional objectives of a unit when stated must be specified in terms of expected learnings outcomes. They must be stated at such a level of generality or specificity that they should neither be too global to be tested nor so elemental or specific as to make an unwieldy list. At the unit level which of the following list of objectives would you prefer to adopt?

(a) The pupil understands the different parts of speech in English.
(b) The pupil understands the use of different types of adjectives.
(c) The pupil cites examples of adjectives of quality.
(d) The pupil understands the meaning of the word adjective.

Since every objective has two major dimensions viz. the ability part and the content part, both must be reflected in its statement. The content part should emphasise more the major ideas, major concepts or principles and not merely the factual information. This is necessary because teaching and testing are both to be geared to such objectives. Which of the following intended outcomes of learning will be more suitable from this point of view?

(a) The pupil recalls properties of oxygen and hydrogen.
(b) The pupil compares the chemical properties of oxygen and hydrogen.
(c) The pupil develops skill in observation of plants.
(d) The pupil develops skill in handling clinical thermometer.

Some objectives are very important because they are prerequisites to other instructional goals. Such objectives play a more significant role in the elementary classes and may be termed as process objectives. Observing, measuring, classifying, using numbers, communicating etc. are process objectives which can be stressed more in the elementary classes. Such

objectives later on act as a means to product objectives like those of knowledge, understanding, application, analysis, synthesis and evaluation. Some objectives may reflect minimum level goals while others may go beyond the minimum level. An objective relating to the recall of certain biological facts may be kept at a minimum level, say, of 50% attainment but an objective relating to the development of skill in handling a clinical thermometer cannot be stated at a minimum level but beyond it, perhaps at not less than 90% level or so.

The ability dimensions of each objective can be well defined (1) into an almost fixed number of specific outcomes of learning in the form of competencies involving pupils' ability to identify, translate, illustrate, relate, compare, classify, interpret, explain, analyse, hypothesise etc. However, the content dimension is unwieldy and has to be reflected in the form of major themes, concepts, principles and processes. This serves two purposes: First, we tend to de-emphasise factual details and stress more the concepts and principles which really form the substantive structure of the subject. Secondly, it is more handy to state the content aspect in a few words.

1.1. *Who Formulates Specific Objectives*

Unlike instructional objectives of the subject which are formulated by a cooperative enterprise of curriculum specialists, subject experts and teachers, the specific intended outcomes of a unit are to be identified or listed by the teacher-evaluator who is concerned with both the teaching and the evaluation of that unit of teaching. In the integral model of evaluation, it is the teacher who specifies the intended outcomes of learning and it is he who evaluates his students on those very outcomes. But in case of a summative evaluation, when an external teacher or an agency is entrusted with evaluation of students learning on the whole or a part of the syllabus, as is the case in a term test, semester or annual examination, these objectives are already prescribed in the curriculum. An evaluator has only to take cognizance of those prescribed objectives. The first step is, therefore, dependent on the purpose of evaluation and the use to which results are to be put. Nevertheless,

specification of intended outcomes of learning is a must, whosoever undertakes its formulation (2).

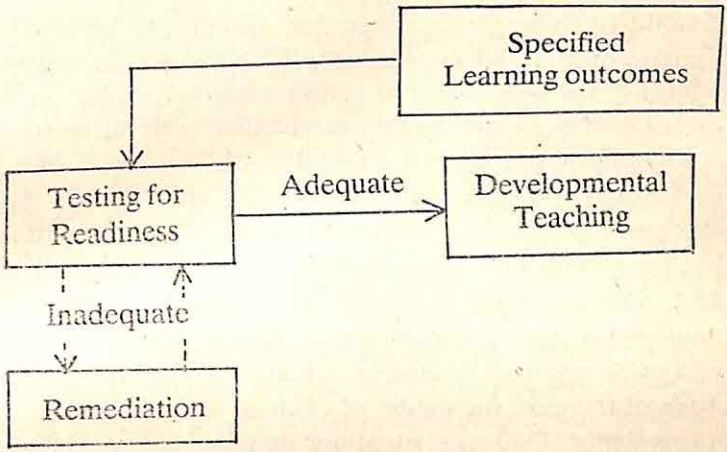
It may be realised that the intended objectives are only hypothesised and all of them may not be achieved if pitched high. It is only after evaluation that we know the validity of these logically formulated objectives. Emergent objectives may not completely be in agreement with the stated (hypothesised) objectives. Emergence of some new objectives in the form of concomitant learning cannot be ruled out.

Anyway, it is these specific objectives which become the reference point for validation of evidence that accrues during the process of evaluation.

Step 2. Appraisal of pre-requisite learning

Since evaluation is treated as integral part of the teaching-learning process, the first phase of evaluation is concerned with testing for readiness so that it may be judged whether students have already acquired the minimum pre-requisite knowledge and skills in order to benefit maximally from the instruction that precedes evaluation of the instructional impact. In other words the objectives specified for the unit can only be achieved if the concepts, skills and information basic to the intended outcomes of learning are learnt. One cannot teach about 'adjective' without the knowledge of 'Noun', multiplication without addition and the mechanism of cellular respiration without the cell structure. Evaluation at this pre-instructional stage is diagnostic in nature to know whether there is adequacy in previous learning or not. Developmental teaching cannot be affective without this. Thus depending upon the facilities and the time available, one can resort to oral or written testing. If there is adequate learning, instruction (developmental teaching) can be undertaken. If there is inadequacy in learning remedial action is called for. It may take the form of reteaching, additional assignments, explaining, discussion, or any other form.

After remediation (if needed) the students are retested for their readiness to know if adequate learning has taken place to take up developmental teaching; otherwise remedial work continues.



In actual practice, a few questions relevant to the intended learning, are put to test pupils for concepts and skills regarded as essential for learning of new concepts or skills. Such questions are put to test their fundamental knowledge and understanding. These questions are used only to explore the possibility of remedial work, if needed, for effectiveness of instruction. It is during developmental teaching that testing in the form of questioning, assignments, hometask, pupils' questions and counter questions is truly integrated. The purpose of this inbuilt evaluation is to improve students' learning by adapting, adjusting and modifying teaching-learning strategies for achieving the desired or intended product of learning. This sort of evaluation is formative in nature as its focus is on adjustment of ends and means to improve student's learning.

Developmental teaching is followed by post-instructional testing, the function of which is certification of students' achievement and their classification for other instructional and administrative decisions. This type of evaluation is summative in nature. Of course, the results of summative evaluation can be used for diagnostic and formative evaluation as all the three types of evaluation are interdependent and have the continuity of a cyclic process. Summative testing, therefore, can be undertaken after each unit, but only for feedback and applying unit correctives (3).

Step 3. Devising situation for testing

Once the intended outcomes of learning are clearly defined or taken cognizance of and the adequacy of pre-requisite learning is established, the next step is to collect evidences about pupils' learning. For this, different types of situations will have to be devised to obtain the desired evidence. As we have already emphasised the unit approach to teaching and testing, these evidences will have to be collected in terms of pupils' attainment of the intended learning outcomes already formulated for the unit. Different types of situations will have to be identified, depending upon the types of evidences to be collected. Selection of a particular situation is closely related to the instructional objectives of the unit, the nature of content and of course, on its practicability. Different situations may be used in different contexts and sometimes in combination of twos or more. Situations may take the form of certain activities, like, reading aloud, projects, assignments, pupils' products, library visits, practical work, a debate, discussion, demonstration, a test, or a quiz etc. Whatever situation may be devised or selected, it must yield evidence which is valid and dependable (4).

Every situation involves some sort of activity which may be used for teaching in pursuance of a particular objective. The same or similar activity can be used for appraisal of the intended learning outcomes also. For example, a simple activity such as drawing a diagram can be used to develop drawing skills among the students and at the same time it may be used as a basis for assessing students' drawing skills. Similarly, many other types of activities can be made the basis of assessment of students' learning. In fact, it is a very natural way of assessing students and may be appreciated especially at the elementary stage where students can be relieved of the fear of examinations that corrodes their mental health.

Projects can also be made the basis of an appraisal of students' learning. Of course, here better opportunities are provided to the evaluators to assess not only students' intellectual abilities but also their skills, attitudes and interests. Outcomes like those of initiative, cooperativeness, social adjustment, sense of responsibility etc. can be more genuinely evaluated using the project as a situation. Nevertheless, in case of individual

projects one can judge more precisely the contribution of each student in terms of his attainment on cognitive, affective and psychomotor objectives. Likewise, an assignment may also be used for assessment of all those cognitive learning outcomes for which instruction was provided. Besides, it provides indicators about one's diligence, systematic organisation, neatness, originality etc. Library visits indirectly show one's interest in extra reading. An excursion or a field trip organised by a Biology teacher provides ample opportunities for the teachers to evaluate pupils' observational skills, collecting skills and cooperative attitude, besides their cognitive learning.

A testing situation is by far the most acceptable and usable form of situation in which a group of students can be examined more validly and reliably in the area of cognitive learning. However, the role of tests in judging students' interest and attitudes cannot be overlooked. The only limitation in this case is the lack of technical competence on the part of teachers and the difficulty of interpretation of test scores. Such test situations may take the form of a written unit test, an oral unit test or a practical exercise, based on the objectives of the unit formulated during the first stage.

Pupils' products which may take the form of a chart, a composition, a drawing or an improvised apparatus can also be used as a situation for judging students' learning. Depending upon the type of pupils' products, outcomes like intellectual abilities, skills or certain aspects of personality can be appraised. Discussions, debates and symposia can be used as situations for assessment of students' ability to express, organise, argue, analyse, criticise etc.

Step 4. Obtaining needed evidences

After specifying the intended learning outcomes and choosing the relevant situations, the next step is to obtain evidences about pupils' learning. This involves the following sub-stages on the lives suggested by TenBrink (5).

(a) To locate information already available.
(b) To decide when and how to get information.
(c) To construct or select the relevant instruments of

evaluation.

(d) To obtain information by using various tools and techniques.

(a) To locate information already available, all relevant sources should be tapped. Teachers' records are the most usable ones for this purpose. It can be a student's performance on previous unit tests, his grades on school assignments or his attendance in various curricular or cocurricular activities. School results may be used for identification data and for longitudinal assessments over a period reflecting students' cumulative achievement. A councellor's record may be used for collecting information about their mental health, vocational interests, attitudes etc. Likewise parents' record if any, can be profitably used to get infomation about some personality aspects like aggressiveness, emotional adjustment etc. Study habits, work habits and health habits are other areas which can be explored from such records. Other records like the ones maintained by a sports' in-charge can also be made use of for appraisal of physical health.

What is it that one should look for in such information? The most important is the *relevance* of information to the desired purpose. *Accuracy* of information should also be checked in terms of errors and the estimate made of such errors if available. How far the information can be useful in making anticipated judgements or decisions? For this cognizance may be taken of the appropriateness of the referents used, evidence of the estimated ability, the evidence of predictive validity and the statement of the information obtained.

(b) When and how to gather information would depend upon the type of evidence required and the use to which this information is going to be put. There are four major techniques of gathering data viz. testing, observation, inquiry and analysis as categorised by TenBrink (1975). Under each of these techniques various types of instruments can be used. Selection of instruments is to be followed by timing of the process of gathering information, specification of testing conditions and finally the construction or selection of information gathering instruments.

(c) If the instruments of evaluation are available with the teachers or the school counsellor, the only task is to select the needed tool and apply it to get the evidence desired. However, care has to be taken that the instrument selected must yield evidence that is valid in terms of the purpose of the test, besides having a reasonable degree of reliability. Depending upon the purpose, greater importance may be given to the validity if improvement of achievement is the focus of measurement. If however, grading or comparison is the aim then reliability may be given precedence over validity of the instrument. In case new tools of evaluation are to be prepared, then all those factors that affect the validity or reliability of the measuring instrument must be taken into account. Ensuring reasonable reliability and validity is the most essential feature of every tool of measurement and forms the scientific basis of collecting information about pupils' growth.

(d) To obtain information, tests or other tools and techniques are to be administered to collect data. Good timing is a prerequisite for obtaining useful information. Are students ready to cooperate for the test or other techniques you are going to apply? They must be prepared cognitively. They must get opportunity to learn from such an attempt. A few unstructured questions, testing rote memorisation of factual information, do not really provide any clue to the evaluator to discriminate between the good and the poor. Any judgement, therefore, made in terms of class performance becomes deceptive. Besides, students are expected to know what it is that they are being evaluated for? In other words if students are given the instructional goals carefully written in observable terms they are likely to learn more from such an experience. Students should also be emotionally prepared to take the test or undergo any other examination. They should view the examination not just as device for labelling them as passes or failures but more as a device to promote their interest and to improve their achievement. This might lead to the development of positive attitudes towards testing and other data gathering techniques.

Teachers themselves should also be sure that they are ready for the test administration. It is not uncommon to find that when a standardised or a newly constructed instrument is to be used, teachers sometimes overlook certain instructions, forget

needed material, ignore time-limits of a sub-test and uniformity of instructions. It must be ensured that any extra material needed in the form of graph, pencils, answer sheets or stop watch must be kept handy and ready.

Step 5. Analysis and recording of information

Information obtained from various sources must be recorded in a meaningful manner so that it could be analysed properly. It is assumed that the information gathered is valid, reliable and up-to-date. What is now required is how to use this information. Depending upon the use to which it can be put an analysis has to be made. Some of the relevant questions that can be asked in this connection are:

(a) Who obtained information about whom?
(b) What is the information and when was it obtained?
(c) How was the information obtained and how far is this information valid and reliable?
(d) In what form (raw score or derived score etc.) is the information available and which instructional objectives does this cover?

Information, if gathered by a class teacher on a unit test, is more reliable for judging his students' learning than about his health habits – about which his parents can provide more accurate and dependable evidence. Similarly, information gathered on a unit test two months ago has no relevance to his achievement now, on another unit test, from the point of view of the use of this information for judgement making. Moreover, the validity and reliability of the information depends on the relevance of the instrument of evaluation used and the reliability of the scores that accrue from it. Once the information is available, it has to be processed into a meaningful pattern so that it could be interpreted in a desired manner.

Information may be in the form of observation notes, test scores, quiz scores, percentage of attendance, letter grades on an assignment, anecdotal record on some significant behavioural aspect, sociometric data, performance in a co-curricular activity or any other data collected through various inquiry techniques

like questionnaire, interview rating scales etc.

Whatever be the data, these must be recorded in a manner that lends itself to easy manipulation. For test data, for example, some code or serial number of the test, scheduled data, test description, testing conditions (context) summary statistics and comments, if any, may be provided on a sheet which becomes a part of the file. For observational data, some format may be desired to record notes of observation, the context in which observation was made, limitations if any and the interpretation made. For recording of data gathered from inquiry techniques, some usable proforma may be developed to note the evidence that accrues from rating scales checklists, questionnaire, interview etc. This should reflect the basis or aspects of measurement, criteria used, mode of summarisation etc. This must be common to the assessment mode used by all teachers. Data on analysis of a product like a composition, an improvised apparatus, an album, a chart or a model can be recorded on the basis of clear description of what was measured or what was expected in the form of a standard or criterion. The rules of assigning marks or grades, the basis of comparison of performance etc. are also to be considered for recording purpose.

Whatever be the mode of recording it must be handy and practicable. Unwieldy records are difficult to maintain and use by teachers and are not much useful to students either. A good record must be simple, up-to-date and readable. Its security as a confidential device may be maintained. Keeping in view the time limitation of the teachers, the procedure and proforma of records should be so simplified that it is easy for teachers to tabulate, collate, maintain up-to-date and use such records efficiently. Further analysis depends on this recorded data. Maintenance of records is a management skill which every teacher should acquire.

5.1 *Three Modes of Analysis*

Collection of evidences is followed by analysis. In fact, analysis begins with the tabulation of data. The type of analysis depends upon the use to which the data are put. Sometimes we are

interested in grading, sometimes in the classification of students, sometimes to promote or certify their achievement. We may be interested more in providing remedial work to the students but sometimes we are more interested in judging our own instructional efficiency in terms of instructional objectives. Perhaps, above all, we are more interested in improving students' learning. In such cases, we want to analyse the evidences to identify inadequacies in students' learning, shortcomings in instructional efforts or pitfalls in evaluation itself. Analysis, therefore, may be done in terms of an individual or it can be done to judge his performance in terms of the class. It can also be done in terms of instructional objectives or any standard of performance may be taken as criterion.

(a) *Self-referenced*

This type of analysis depends on judging the performance of an individual in terms of his own abilities and efforts. Since every individual learns at his own pace and has a varied educational, social and experiential background, his performance cannot and need not be judged in terms of the class as a whole. Self-referenced analysis is done with reference to one's own performance or progress. As such it provides better feedback for his past achievement and is a good indicator of his future achievement.

(b) *Norm-referenced*

Students' achievement can also be judged with reference to his class in the same section and even with reference to a school, a district or a board of school education or the state. Here we are more interested in judging the performance of an individual against a norm or average performance of the group. This is at present the practice in most examinations conducted inter ally or externally because the purpose of such examinations is to classify students for promotion or to grade them into different categories such as first, second or third divisioners. As far as learning is concerned, the comparison of a student's performance with the group is not as significant as the knowledge about the inadequacies in his learning. Comparison of students' performance in terms of a norm or a group is useful when we

are interested in the comparison of institutions or the selection of students.

(c) *Criterion-referenced*

Another way of looking at the performance of a student is to compare his performance with reference to some pre-determined criteria e.g. instructional outcomes. Intended instructional outcomes may be visualised in terms of an instructional objective of a unit, a particular level of proficiency in a skill or a minimum mastery-level of achievement. Unlike norm-referenced comparison in which we may be satisfied with four correct spellings out of 10, the criterion-referenced approach emphasises that a minimum of 8 or 9 out of 10 spellings can be considered a desirable level of performance. If we are interested in bringing up our students to a particular level of achievement to enable them to derive maximum benefit from further instruction, this approach should be adopted. It envisages analysis of evidences in terms of a pre-determined, acceptable level or standard of performance. This approach envisages that every individual is supposed to reach almost the same intended level of achievement or performance although some may take more time than others to reach that level. Thus analysis may be done in terms of the individual, with reference to the criterion of performance.

Step 6. Forming judgements and taking decisions

Evaluation is a process of making value judgements. Therefore, analysis of evidences is followed by interpretation of the evidences to form judgements. There are various types of judgements which the teachers form. Grading of students, diagnosing their weaknesses, predicting success in the examinations etc. are day-to-day judgements that the teachers make. Based on such judgements they take decisions regarding promotion, remedial work or changes in instructional strategies.

Interpretation depends on the conversion of scores, grades, or any other type of evidences into more meaningful terms so that the students and parents can understand it. For this, comparison of evidences in terms of some referent is essential. Such a referent may be an instructional objective, a class or the

individual himself. Accordingly, we may have three bases of judgement: self-referenced judgements, norm-referenced judgements and criterion-referenced judgements, as already mentioned in the previous pages. Self-referenced judgements are based on comparison of performance in terms of one's own rate of progress while norm-referenced judgements involve comparison of a student's performance in terms of the group, a class, district, state or national level. Criterion-referenced judgements on the other hand, refer to some sort of criterion which may be an instructional objective, a performance standard or a learning outcome. These three bases of judgements form one dimension of judgement-making.

The second dimension of judgement-making concerns the type of judgements. For example, an individual's performance can be judged in terms of his deviation from the norm, like that of a class performance or school performance. It means the extent to which an individual deviates from the average performance of a group to which he belongs. Such judgements which emphasise one's position in terms of deviation are called deviation judgements. All our classroom judgements and Board results are examples of this type. The other mode of forming judgements can be in terms of mastery learning. According to this approach, a student is to be judged in terms of the mastery of what is learnt. In such cases it is expected that the student acquires mastery beyond 75 or 80% level of the intended level of performance or the pre-determined criterion of expected learning outcomes (6). Therefore, 8 correct spellings out of say, 10, might be considered as a minimum level of mastery in one case while the development of 10 out of 10 component skills needed to use a clinical thermometer efficiently may be considered essential for mastery in another case.

The third dimension refers to the purpose of judgement-making. Accordingly, judgements may be made to estimate one's present status, position or achievement. It may a predictive judgement, where the purpose is to make prediction. What is needed is the selection of the predictor, the collectism of evidences on the predicted variables and the estimate of the relationship between two sets of data. Therefore, judgement forming is closely related to the referent selected and the mode

of interpretation. In day-to-day classroom assessment on teachers' assignment and tests, values are assigned in the form of grades as A, B, C, D, E, etc. Such judgements are made sometimes according to normal curve or as percentages of grades, scores, level of achievement or grading on a standard (mastery) judgement or self-referenced judgements.

6.1. *Making Decisions*

Learning to make judgements for a better position is essential for a class teacher because it is on the basis of these judgements that decisions are taken. In fact, decision-making is the ultimate aim of evaluation. Faulty judgements can lead to wrong decisions. It is, therefore, essential that judgements should be verified first and only then decisions arrived at. The accuracy of information, may be checked to verify judgements. For this, relevance of data gathering instruments and their reliability are two important aspects to be checked. Judgement-making becomes useless if the error of measurement is appreciable. Further empirical evidences may be collected in such cases or be prepared for review of a decision.

Decision-making is a complicated process and requires a critical approach to avoid situations which later on demand reversal of decisions made earlier. Faulty decisions may also be due to imprecise statement of instructional objectives, taking cognizance of only few alternative solutions, failure to visualise each alternative possible use of invalid judgements and overlooking professional ethics. Therefore, whenever decisions are to be taken about grading, certification, remediation, change in instructional stretagy, etc. teachers should be cognizant about the following steps:

(a) Specify your objective clearly in terms of expected or desired learning outcomes.
(b) Identify all possible alternatives and probable outcomes of each alternative.
(c) Visualise the likely consequence of each of the courses of action visualised, along with probable outcomes.
(d) Choose the best alternative by giving appropriate value to

different alternatives and weightages to probable consequences.

Judgement therefore, is a very responsible job and requires critical acumen besides the technical competence required to process the data in terms of pre-determined purposes. Faulty judgements are the result of unreliable and invalid evidences, or due to incognizance of the purposes for which the judgements are made. Like-wise wrong judgements lead to wrong decisions; thereby wasting all efforts used in collecting relevant evidences about pupils' learning.

Step 7. Summarising and reporting

Once the judgements are formed and decisions are made, two possibilities emerge. If everything is alright, summarisation and reporting follows. If there are certain inadequacies in the achievement of students then diagnosis and remediation may be taken up before summarisation. Summarisation involves the types of evaluation reports made to communicate to its users. Learning the technique of summarising the results and application of specific rules to communicate the results to those who are interested in the improvement of pupils' learning are, therefore, essential features of reporting.

Single numerical scores comprising the subject-wise data which is in vogue, smacks of an authoritarian approach and a bureaucratic convenience of recording, filing and storing records in the past. Complete exclusion of data gathering devices other than testing was, perhaps the main reason in the past why teachers' assessment were outweighed by his Majesty's Inspectors. Now with the use of a variety of evidences of different aspects of students' growth, summarisation and recording have become more cumbersome and technical.

Properly summarised evaluation data provides the basis for making further judgements and decisions by administrators, parents, teachers and other users. Good summaries are good indicators for discovering the cause of unplanned instructional efforts at a particular time. Moreover, they are really used for judgements made or decisions taken.

The first requisite for preparing a good summary is to

identify the clientele. Is it for the parents, teachers, students, parent-teacher meetings or for some other purpose? The summary should indeed be made by the teachers and for the teachers as these are to be used, by and large, by them to improve students' learning and their own teaching. What is the purpose of the summary or for what summary is to be used, determines the types of report to be prepared. A letter, a report card, a diagnostic report, a conference, an action report etc., are different techniques for summarisation. The mode of summarisation depends on the use to which the report is to be put. It may concern placement in a remedial class, shortage of attendance, an award for the best student, the timing of completion or the beginning of a unit, etc.

The next step in reporting is the listing of the decisions made. Only the critical decisions should be included, in order, from the most recent one to the remotest one. Each decision conveyed must be supported by the judgements of the teachers who influenced those decisions. Lastly, mention must also be made of the information that led to a particular judgement to indicate the basis of judgement-making. Thus summarisation includes critical decisions based on judgements formed and supported by information that led to those decisions.

Summarisation is a precursor to reporting. A report means communicating the summarised evaluation data to the users. It may be formal or informal, written or oral, pre-planned or emergent. A very clear, unambiguous and easy-to-understand report only can be appreciated by its users. It must include all the minimum information essential to the intended users. If a statement of what is being reported clearly indicates whether it is about the behaviour, class performance or a parents' meeting it becomes more meaningful Guidelines must be provided for interpretation of the information symbols. Each report should be meaningful and actionable for the recipient. For example to say, "Your ward is not getting on well", is meaningless for the parents. But to say, "He is weak in English spellings when compared to his classmates", is more meaningful because the parents can take some action in the form of giving dictation or spelling tests at home. Reporting is, therefore, the service-component of the evaluation process on the basis of which a

relationship among teachers, students and parents is established.

7.1. *Diagnosis and Correctives*

As pointed out earlier, summarisation and reporting follow a decision strategy, if everything goes well in terms of students' learning. However, if certain inadequacies in students' learning come to the surface than further diagnosis, followed by remedial work, precedes summarisation and reporting. This step indeed refers to the pedagogical aspect rather than the measurement aspect of evaluation. Diagnosis may be focussed on the teaching-learning process or more often on the pupils. Diagnosis followed by various types of correctives should in fact be a regular feature of any formative tests. Diagnosis is, therefore, of necessity imposed by the judgements from the evaluation data. Diagnosis, indeed, should be considered a sacred service of the teaching-learning process (7) and thus becomes integral to the instruction. Various types of correctives can be applied, depending upon the nature of difficulties diagnosed.

Correctives no doubt relate to the padagogical aspect as their focus is on the improvement of instruction vis-a-vis students' learning. There are various types of correctives like alternative textbooks, workbooks, flash cards, re-teaching, audiovisual material, programme instructions, tutoring etc., which can be applied (8). Such correctives can be individual-based or group-based depending upon the nature and magnitude of errors observed. Some of these correctives can be used either for the whole class, for a smaller group or for an individual. These correctives can be involvemental correctives, if a learner is himself involved in the learning process e.g. self-study session. Sometimes presentational correctives can be used which are good for a group-based plan of instruction. They focus primarily on presenting the new material like the re-teaching of a module. The choice of a corrective is, therefore, of primary importance for improving students' learning. Further details of diagnosing pupils' weaknesses and the related remedial measures will be discussed more thorougly in a separate chapter on diagnosis and remediation.

Step 8. Feedback and correctives

Since the focus of evaluation is on the improvement of students' learning and instruction, the evaluation process cannot be considered complete unless the evidence is made use of for this purpose. Feedback mechanism provides loops to each and every component of the teaching-learning process. As such feedback can be made use of in the following manner (9):

(a) *To judge the efficacy of the objectives*

Test results can be examined in terms of the objectives set up. Constant poor results in relation to a particular objectives, say the application objective, may indicate that the objective may perhaps be pitched too high for the students and needs modification. Other things remaining the same, if most of the students do exceptionally well on a test, it might reflect the lower level of the objective which may be suitable for a lower grade. Similarly, if the results are analysed in terms of specific learning outcomes (behavioural outcomes) implied by each objective, it can also serve as a basis for modification, deletion or addition of certain outcomes of learning depending upon the level of students' performance on each.

(b) *To improve curriculum*

Analysis can also throw light on the suitability or otherwise of the curriculum as well. Too difficult or too easy areas of content may lead to modification or even their total exclusion from tne course. Over-emphasis or under-emphasis on a particular topic, misplacement of a concept, sequencing and reorganisation of certain units, proper grade placement of certain concepts etc. may require further investigation. All these improvements relate to curriculum redevelopment.

(c) *To promote learning*

Assessment of pupils' performance helps to understand the level and learning of various concepts. Immediate feedback of results to students, teachers, and if needed to the parents as well, also goes a long way in improvement of their achievement and reinforcement of what is learnt. Moreover, knowledge of inadequacies provides them direction for further improvement. The most significant contribution that results from such a

feedback is the development of a positive self-concept among the pupils. Continuous feedback of the pupils' adequacies acts as a great motivating force for further learning.

(d) *To improve instruction*

Results of measurement indicate the emphases laid in teaching. The quality of teaching is reflected indirectly in pupils' learning. Therefore, on the basis of students' performance, the teacher can rectify his past mistake, and prevent their recurrence in future. The performance level of the students on various instructional objectives, indirectly shows success, failure or limitation of methods and materials used in instruction. Likewise, contentwise analysis of performance may also indicate whether a unit or a module has been learnt well, failing which, it might require re-planning and re-organisation of activities for future instruction.

(e) *To provide guidance*

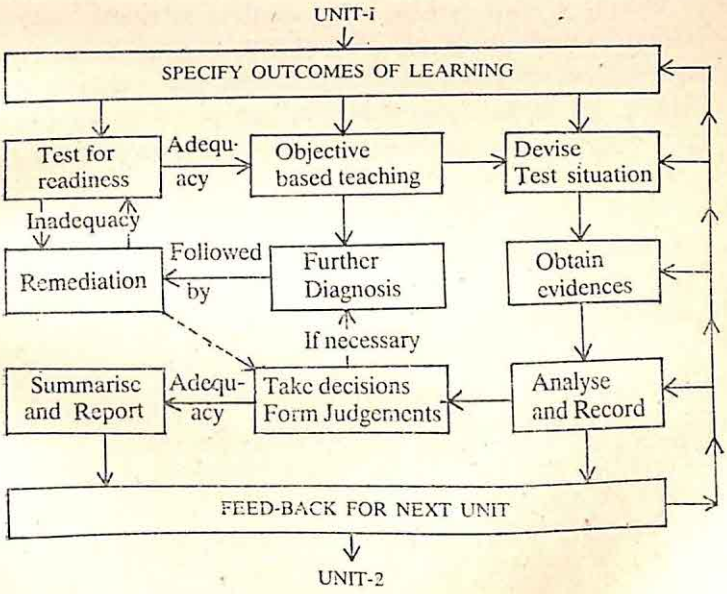
Guidance without measurement loses it scientific character while measurement without guidance loses its purpose to a great extent. Thus, results indicate the areas of successes and failures on the basis of which future success can be predicted. Poor performers are identified and their weaknesses are located, so that necessary guidance could be provided. Study habits are conditioned by methods of examinations. On the basis of scores, students can be guided in the selection of right courses to help them make right vocational choices later. Provision for individualised instruction for certain students and enrichment programmes for other individuals can also be planned and organised.

(f) *To promote research*

Action research is the means of finding out the truth regarding questions of practical importance concerning the job and not research dealing with theoretical questions or eternal truths. Advantage can be taken of the analysed data by using them to solve certain problems like the impact of teaching methods, choices of questions in a question paper, types of options, allotment of marks for individual questions etc. Problems like

these can be solved by undertaking classroom research and using the data for improvement of learning and instruction. Once the feedback is done, the process of evaluation with respect to a unit of teaching comes to an end. Then follows the next unit of teaching.

In this integrated model of evolution, one may appreciate that evaluation of students ends neither end with the collection of measurement data nor with analysis and judgement-making functions but goes beyond that to feedback the results to all the components of the teaching-learning process. It is this use of results of evaluation that helps the teacher to reformulate instructional objectives, provides a basis for reorganisation of learning experiences, identifies teaching-learning defficiencies and even judges the effectiveness of the evaluation programme itself. One thing which needs to be appreciated is that in formative evaluation, an attempt is made to integrate teaching, learning and testing. Teaching provides the validating referent while testing becomes an inseparable part of the teaching-learning process. Therefore, the use of evaluation as a device for furtherence of instructional efficiency as well as for improvement of students' learning should be considered as the ultimate aim



of teacher-evaluator.

The relationship among the various steps of evaluation discussed above can be shown in the form of an integrated model of formative evaluation as shown on page 185 (10).

CHAPTER VII

TOOLS AND TECHNIQUES OF EVALUATION

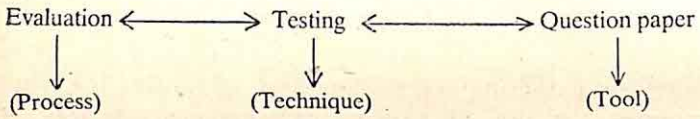
1. Introduction

Information gathering is the first step of evaluating students. It requires the use of various data gathering devices. These include different types of tools and techniques which are applied to collect various types of evidences. A number of tools are employed relating to various techniques like those of testing, observation, inquiry etc. What is important is the selection of the right device for collecting the needed information and the quality of data that accrue from the application of those tools and techniques. Attempt will be made to give here brief description of these devices and their role in data gathering. It is expected that readers after going through this chapter will be able to

(a) describe the nature, scope and limitations of various tools and techniques,
(b) classify the various tools under different techniques of evaluation,
(c) identify the relationship between the types of evidences possible through application of techniques like testing, observation, inquiry and analysis,
(d) select and apply the appropriate tools and techniques to collect the needed evidences about pupils' learning.

2. Testing Techniques

Technique refers to the mode of collecting evidences about pupils' growth or achievement whereas a tool is an implement or instrument that helps us to extend our observation by bringing more precision in collection of information or data. Thus a technique involves the use of certain tools for data collection or summarisation of data. The four basic techniques of evaluation are: testing, observation, inquiry and analysis. Each of these techniques when applied involves the use of various tools. The following flow chart would make the relationship between the two more clear:



There are two types of evaluation devices, namely, the data gathering and data summarisation. For data collection these may be testing techniques or non-testing techniques like observation, inquiry and analysis. With each technique a number of tools are associated. Some of the more commonly usable tools are categorised and listed under of the testing and non-testing techniques in the form of a flow chart (see annexure) followed by a discussion of the more commonly used technique.

2.1. *Written Examinations*

Teacher-made tests could be written, oral or practical. It is the written tests which have been considered more reliable for measuring learning outcomes in the cognitive domain as also for the affecting domain. In the cognitive domain, written tests are in vogue in all subjects. Their prominent use in evaluating students at all levels is due to their better reliability than the other two types. Inspite of the fact that the present written examinations suffer from many defects, they are still the most acceptable device for judging students' achievement and even

for selection of students for various admission tests. This does not, however, mean that written tests as they are now constructed, administered, scored and used are a quite reliable and valid measure of students' achievement. They need to be planned more scientifically, developed more technically and interpreted more cautiously.

The written tests more commonly used are the achievement tests but other varieties like diagnostic tests, criterion-referenced tests and selection tests are also in use besides attitudes, scales, interest inventories etc. For details of criterion-referenced tests and diagnostic tests separate discussion is required. For planning and development of achievement tests, a more detailed discussion is warranted as teachers, paper-setters and evaluators are all concerned with the use of such tests. As such the next chapter is wholly devoted to the planning and development of achievement tests. Thus more attention is given here to oral testing and non-testing techniques viz. observation, inquiry and analysis.

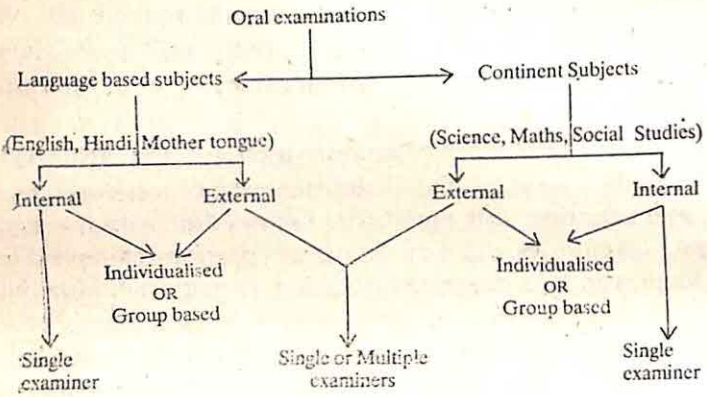
2.2. *Oral Examinations*

Oral Examinations are perhaps the oldest form of evaluating students prevalent in most countries till the beginning of the 18th Century. Till the invention of the printing press, the oral word used to be the only means of communication as also the only medium of testing students.

Although in countries like the USSR, oral examinations have still a permanent place yet the fact remains that all over the world written examinations are so much in vogue that they have almost relegated oral testing to the background. One major reason for it is the ease of construction, administration, scoring and interpretation of written examinations and the possibility of maintaining better validity and reliability of expected responses as compared to oral examinations. (1)

Nevertheless, oral examinations have their own place and cannot be replaced by written examinations. Oral examinations demand oral evidence, although questioning can be oral or written. Of course, recording of evidence has to be in writing except in cases where the group of examiners is extremely

small. Oral examinations may be classified as under:



(i) *Purpose*

The purpose of oral examinations is fourfold:

(a) To test oral skills like listening, speaking, reading and expression (oral) which cannot be tested through written examinations.

(b) To confirm and probe further the extent of evidence gathered through written devices whenever desired, as in a project, a unit test or a viva voce.

(c) To judge the extent to which such skills as are warranted by the nature of the subject, are developed, like argumentation in history or observational skills in biology.

(d) To make a quick oral review for informal assessment of what the pupils have learnt and what their deficiencies are.

Thus an oral examination may be used to test content as well as manner. Its advantage is that oral skills, work experience, project work, functional understanding about practical skills etc. can be validated more profitably through oral testing which act as an external moderation device. For younger children, it is an indispensable testing device which provides on-the-spot feedback about students' learning and instructional strategies, thereby providing a basis for remediation. Likewise, students weak in written expression can be assessed better

through oral devices.

In the fifth conference of the Chairmen and Secretaries (2) of the Boards of Secondary Education, held at Azad Bhawan, New Delhi, in 1964, the then Education Minister, Sh. M.C. Chagla, stressed the need for developing oral expression which is very essential in day-to-day living. This ability is desired in every walk of life and every interview one faces for employment. Despite realisation of the need for the development of oral skills, the fact remains that unless these skills are evaluated in external examinations they are not going to attract the needed attention of teachers in developing these skills. Besides, the subjectivity in assessment, the need for greater number of examiners and their specialised skills in testing say, pronunciation; inter-examiner variance, its time-consuming nature, the difficulty of the objective recording of evidences, ensuring equivalance of questions, subjective interpretation etc. are a few among the many more problems and difficulties in using oral examinations at the external level. However, as an informal assessment device, oral questions in all subject areas would continue to be taken advantage of by class teachers for diagnosis and remedial programmes.

(ii) *Improving the validity and reliability of oral examinations*

Since by and large oral examinations are to be used in the field of languages, it is necessary that for systematic evaluation the following procedure may be tried:

(a) Identify the objectives (listening-speaking-reading, expression)

(b) Delineate the aspects of evaluation (say, oral expression)

In oral expression, aspects like language and content elements may be identified.

(c) Define oral expression as under:

The pupil

(i) uses appropriate words in the right context.

(ii) uses the correct sentence pattern.

(iii) pronounces words correctly.

(iv) modulates his voice according to the situations.

(v) speaks at a reasonable speed.

3. Observational Techniques

Next to testing the most commonly used technique is the teacher's observation. Observation is specific, planned, systematic and recorded by an expert as contrasted with casual seeing. It is a process in which the observer listens, looks at or notices the significant elements of a product or performance of a student which is aimed at collecting information about a person's cognitive, affective or psychomotor behaviour. Certain skills like oral reading, listening, speaking, dancing, singing etc. cannot be measured through paper-and-pencil tests but can be measured through the observational technique. Other examples where this technique can be used are attitudes, interests, habits, leadership quality, social adjustment etc. Such skills which are either themselves observable like singing, dancing and speaking or yield observable products like those of painting, writing or drawing lend themselves better to the use of the technique of observation. It is an excellent technique to obtain evidence about the typical behaviour of students as they take a test, play or behave in a situation. Observation is one of the most convenient techniques which can be used profitably, provided it is properly planned with a specific purpose, systematically carried out, carefully focussed, thoroughly recorded and cautiously interpretted by an expert. (3)

Though this technique is quite subjective in nature, yet the validity and reliability of this technique can be improved if care is taken in the construction and use of the instruments of observation. Frequent observations at regular intervals by the same observer or independent observations by more than one observer at the same time can definitely reduce subjectivity in the assessment of typical behaviours. Moreover, the barrier between the observer and the observed can be removed by making observations in a natural setting, under normal conditions, without making the subject of observation conscious that he is being observed.

3.1. *Instruments of Observations*

For recording of information collected through observation, a

number of devices can be used. The major instruments af observation are checklists, rating scales, rankings, score cards and scaled specimens. Of course tools like, the lense, the microscope, the camera, the stop watch, the thermometer etc. are used to aid observation.

(i) *Checklists*

It is one of the simplest devices, consisting of a prepared list of items which provide information as to whether or not a given characteristic is present. It is just like a laundry list, the presence or absence of which can be indicated or recorded in the form of 'Yes' or 'No'. Half-way responses are not admissible in such a list. The statements are given in specific and functional terms. While preparing the checklist we have to analyse the job, the task, the skills etc. into various components which should be included in the list. Such a checklist facilitates and systematises the recording of observations.

An example of a checklist on study skill

1. Can I read silently?
2. Can I consult the dictionary?
3. Can I read at a reasonable speed?
4. Can I relate ideas in the para?
5. Can I grasp the central idea of the para? etc. etc...

Students are asked to tick-mark the statements which they think they can answer in 'Yes' and cross mark those which they would like to respond in 'No'. The analysis indicates the adequacies and inadequacies in study skills and provides a basis for teachers and students for self-evaluation.

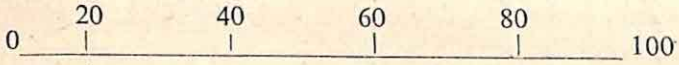
(ii) *Rating scales*

The rating scale involves a qualitative description of the limited number of traits of a person or aspects of a thing in contrast to a checklist which shows merely the presence or absence of a trait. Important aspects are selected and the observer's attention is directed to those specific aspects. This instrument ensures more accuracy and objectivity of judgements to be made by the observer. We can use this tool for observing the behavi-

our, performance or the end products of his performance.

We can have 3-point, 5-point or 7-point rating scales, depending upon the refinement and practicability. (4) In a classroom situation a 3 point or 5 point rating scale is more commonly used and is more usable. Such scales can be:

(a) descriptive, when a person is rated by descriptive phrase like, say, excellent, good, average, fair, poor (5 point)
(b) numerical, when a numeral is assigned to each descriptive phrase 4, 3, 2, 1, 0 (5 point, corresponding to terms like always, frequently, occasionally, rarely, never
(c) graphical, in which case descriptive phrases are represented on a line at various points in the form of a continuum, say, ranging from 0 to 100. The student is to check mark on that line, at the appropriate place, as shown here.



The difficulty of defining a trait or a characteristic to be evaluated, the halo effect of the carrying effect of a judgement from one trait to the other, tendency of raters to be generous in rating, are some of the glaring limitations of rating scales.

(iii) *Anecdotal record*

An anecdotal record is a verbal snap shot of a single but significant event recorded in its setting to describe observations made by the observer (5). These descriptions are a factual record and not interpretations of what happened. It is a 'time lapse recording technique' like a single framework camera. Such records are useful for collecting information about the social adjustments of students on a variety of behaviours which are typical. A teacher can keep on recording such information in these anecdotal record cards and can make use of this information whenever a students' record of typical behaviour is called for. Such records are to be interpreted very cautiously, especially keeping in view the context or setting in which a particular event or behaviour occurred. They form the recorded behaviours of the observer for use in forming judgements about students' behaviours.

(iv) *Rankings*

It is a crude observational tool which is used simply to rank students on the basis of impressions and is therefore quite subjective. The students or the objects of measurements are marked on the basis of the degree to which they possess or do not possess particular characteristics. Ranking may be done on the basis of:

(a) grouping, where predetermined number of groups are formed e.g. a product may be ranked into such groups as 'average', 'above average' and 'below average',

(b) the paired-comparison method, in which case a student is compared to every other student and the rater makes judgement as to which of the two is better. This technique can be used for 'ranking individuals, the objects and the product. It is rough but easy to use technique. Its disadvantage is that the significance of a rank is dependent on the size and nature of the group. For example, a rank 'third' out of a group of 10 is not the same as that out of a class of 40 students. Likewise a 'third' rank out of a group of 25 gifted children and a 'third' rank in a group of 25 below average students carries altogether a different connotation. This technique, being of limited use, should be used only when a quick and rough estimate is needed in which a relatively small group of individuals or objects are involved.

(v) *Score cards*

It provides for appraisal of a larger number of aspects or traits and carries the advantages of both the checklist and rating scales. Like the checklists, different aspects of assessments are identified and listed. Each aspect is weighted by giving a predetermined point value like the rating scale. A numerical value is given to the various rating points, say 4, 3, 2, 1, 0 (5 point rating scale). When any aspect is to be rated, say, review exercises in a textbook, the weighted score for each aspect as well as the total weightage can be used to evaluate the object observed (6):

Example: A review exercise of a textbook, containing 10 questions

Aspects	Weightage	Ratings (0—4)	Aspect-wise weighted score
		0 1 2 3 4	
(a) Coverage of objectives	25	— — √ — —	25×2= 50
(b) Coverage of content	25	— — — √ —	25×3= 75
(c) Preciseness of language	10	— — √ — —	10×2= 20
(d) Clarity of scope	15	√ — — — —	15×1= 15
(e) Appropriateness of directional words	15	√ — — — —	15×1= 15
(f) Difficulty level	10	— — — — √	10×4= 40
	100	Total weighted score	=215

Score cards are normally designed for evaluating communities, schools, textbooks, socio-economic status of families. As with the rating scale, the limitation of score cards is the difficulty of choosing aspects to be appraised, identifying and quantifying them. Moreover, the fact that the whole is greater than its parts, is also a point for consideration.

4. Inquiry Techniques

Next to testing and observation, the most usable technique is the instrument of inquiry which help the teacher to collect information about an individual's perception of himself or of others in a more objective and systematic manner. The questionnaire schedule, socio-metric devices, projective techniques, attitudes scales and inventories are the more commonly used devices.

(i) *Questionnaire*

A questionnaire is a set of written questions listed in a manner that can be read and responded to by teachers, students or other

respondents when mailed. If administered in person, the information gathered through a face-to-face interview it is called a Schedule. (7) A questionnaire is usually duplicated and mailed to a large number of respondents to be filled up and returned. Its focus is on obtaining information about the opinion and attitudes of individuals or even what an individual did or might do in a particular situation. A questionnaire may be the 'closed form' or the 'open form' type. In the former case, the respondent is asked to mark 'Yes' or 'No' or to check an item, while in the latter we call for a free response in the language of the respondent and thus provide for greater depth of responses. Most questionnaires include both forms as indicated below by the three questions to illustrate the three forms:

1. Which of the two vocational courses, medical and engineering do you prefer? (Closed form)
2. Why did you choose a medical course in preference to an engineering course? (Open form)
3. Which of the two courses of studies, medicine or engineering do you like more? Give reason for your choice. (Combination of the two)

Questionnaire are useful tools to collect opinion from various types of respondents, in a comprehensive manner, involving less expenditure. However, the disadvantage is that what the students perceive, say, about the relevance of a syllabus, may not be what actually is. They yield perception of what is and not necessarily what actually is. Nevertheless, the teacher can change the perceptions of the students by changing his methodology of teaching after analysing their perceptions as evidenced in their responses to the questionnaire. Another limitation is that respondents usually do not fill the questionnaire seriously and also try to show themselves better than they actually are or what they feel about the aspects asked for in a questionnaire.

An inventory is a special kind of questionnaire used for getting self-reports and is very highly structured. It provides a checklist of behaviours in the form of likes, dislikes, habits,

opinions etc. An attitude scale is another variety of questionnaire which represents a combination of the rating scale and the self-report questionnaire. The respondent is asked to match his 'feelings' with those described in the scale that is usually given in the form of a continuum ranging from, "strongly agree," to "strongly disagree". The respondent reads the statement and tick marks the one to indicate if he strongly agrees, is undecided, disagrees or strongly disagrees (Commonly used form).

(ii) *Interview*

An interview is a face-to-face conversation with a purpose which can be informal and unstructured or formal and highly structured. In a way, it is an oral questionnaire in which the interviewee gives the needed information verbally in a face to face relationship. A good interview calls for proper scheduling, a natural setting, adequate planning of questions, establishing rapport with the interviewee, lonely conduct of interview, systematic recording of responses and cautious interpretation of responses. An interview provides the teacher with information about interests, attitudes, opinions, self-perception and typical behaviour patterns of students when conducted in a usual routine of interaction between the teacher and his students. An interview helps the teacher to keep track of changing perceptions, attitudes, personal relationships etc.

This technique is useful while dealing with children, illiterates and those with language difficulties besides those students who have personal and social adjustment problems. A deep penetration into the reasons for actions, feelings and attitudes is possible with a skillful interviewer to motivate the learners through this techniques. However, this technique is time consuming and interviewer-biased and needs expertness to ensure objectivity, sensitivity and insight on the part of interviewers. The unique advantage however, of this technique, is that teacher can obtain effective information of a highly personal nature about students.

(iii) *Sociometric instruments*

Sociometry is a technique of describing social relationships

among individuals in a group. Sociometric instruments are designed to collect evidence about the social acceptance of individuals within a group as also about perceived relationships existing within that group. Children can be asked to nominate a child or give names of three or so, in order of preference, concerning say, invitation for an excursion, a monitor for the class, the captain of the hockey team, a desk mate, project-group etc.) It is a peer rating, rather than rating by teachers or superiors. Moreno (8) launched sociometry in 1934. Different types of devices are usable.

(a) *Guess-who technique* (Hartshorne and May, 1928). Children are asked to name an individual who fits a certain description. An instrument describes a number of individuals and the respondents (students) are to guess who is being described. It may be a loner, a leader, class clown, the class brain, the trouble maker, compromiser. Statements like the following may be asked followed by the question, "guess-who"?

1. This boy is always worried.
2. This boy is always picking quarrels with others.
3. This boy is never prepared to work in a group project.
4. This boy likes to be left alone all the time.

(b) *Social distance scale* (E.S. Bogardus). This scale attempts to measure the degree to which an individual or a group of individuals is rejected, or accepted by another individual or a group. Various situations are scaled and score values are established which range from complete acceptance to complete rejection. An individual is asked to check mark his position by selecting one of the points on the scale. For example, a student in a classroom situation can choose one of the following statements do check his position:

1. I would like to have him as my best friend.
(Complete acceptance)
2. I would not mind having him as my class monitor.
(Partial acceptance)
3. I wish he were not in my class. (Rejection)

The scale can be of 5 or 7 points more precise measure of acceptance or rejection.

(c) *Placement devices.* To discover perceived relations, a situation is described pictorially or verbally and the individuals are asked to place the members of the group in different positions which are described. For example a map or diagram may be given, depicting a playground and the students may be asked to put the names of their class fellows in different positions indicating where they would most likely be found during recess.

Although it is difficult to interpret socio-metric data, yet the information gathered is useful for making decisions about grouping of students for various activities. The use of such devices at regular intervals helps a lot in making self-referenced judgements about students' personal social adjustments.

(iv) *Projective techniques*

Projective devices are used to enable a subject to project his inner feelings, needs, values or wishes to an external object. The subject reveals himself unconsciously while reacting to the external objects. They are designed to find out about personality and social adjustment characteristics of individuals. The individual is shown some pictures and told to describe what he sees and his responses are categorised and analysed. When ambiguous stimuli are given, it is assumed that the individual projects his biases, feelings and idiosyncracies. Since these instruments are highly unrealisable, they must be used with the utmost care and caution.

Projection may be sought through:

(a) association in which the individual is to report what he sees, thinks or feels when he is presented a picture, an inkblot, a phrase or word. The Rorschach inkblot test and the Thematic Apperception Test are familiar examples. (9)
(b) completion, in which the respondent is asked to complete a sentence or a task e.g. My greatest ambition in life is....
(c) role playing, in which case the subject is asked to act or improvise a situation according to the role assigned to him. Frustration, insecurity, hostility, prejudice etc. can be

observed by this technique.

(d) *Creative and constructive work* whereby the subject is asked to model clay' paint, play with toys or write imaginative stories on the given topic. Deep-seated feelings may be inferred by noting the choice of colour, form, words, orderliness, evidence of tension etc.

5. Instruments of Analysis

Analysis can be used independently or along with an observational technique. Content-analysis can be used to analyse the students' work as he learns. Written or spoken communication can be analysed for the presence or absence of certain characteristics say, like spelled words or a creatively used verb. These are counted and compared with some standard to make a judgement about the quality of communication. The most important instrument of analysis is the analysis of projects and assignments which specify the important characteristics in advance. Three types of projects and assignments are significant from the point of view of analysis. TenBrink gives the following three types of assignments (10):

(i) *Acquisition assignments*

These assignments are prepared to help students to acquire new information or skills. Following examples are relevant in this regard:

(i) Study unit 5 and identify the new terms introduced for the first time OR study the spelling words for the Monday test.

(ii) Memorise your table of multiplication for 6×10.

(iii) Read chapter 3 to know different modifications of root. For evaluative information many variables that effect learning are to be controlled otherwise it would be difficult for the teacher to find out how information is acquired.

Such assignments help to know what new knowledge and skills have been learnt; how fast did students learn and what kinds of errors they make during acquisition.

(ii) *Review assignments*

In order to reinforce the learned information and skills, review is essential. Recurrent use of concepts in science textbooks or practice in using newly acquired words, reading skills etc. help the learner to retain that information and skills. Well designed review exercises enable the teacher to know the process a student uses for review as also about the level of original acquisition. Analysis of review assignments help the teacher to know about unexpected outcomes also. Ask students to show sums in Mathematics or a composition in English.

(iii) *Transfer exercise*

Knowledge and skills acquired in a subject help a student to know more about himself even in other situations in real life. Transfer exercises should enable students to transfer their learning (knowledge and skills newly acquired) to new learning situations. What are the conditions under which transfer is facilitated or inhibited? They provide a wide variety of situations to try out newly learned knowledge and skills. Transfer exercise must be followed by a review exercise because it is necessary to know what information or skills the students have retained before you can judge the transferability of the various concepts, skills etc.

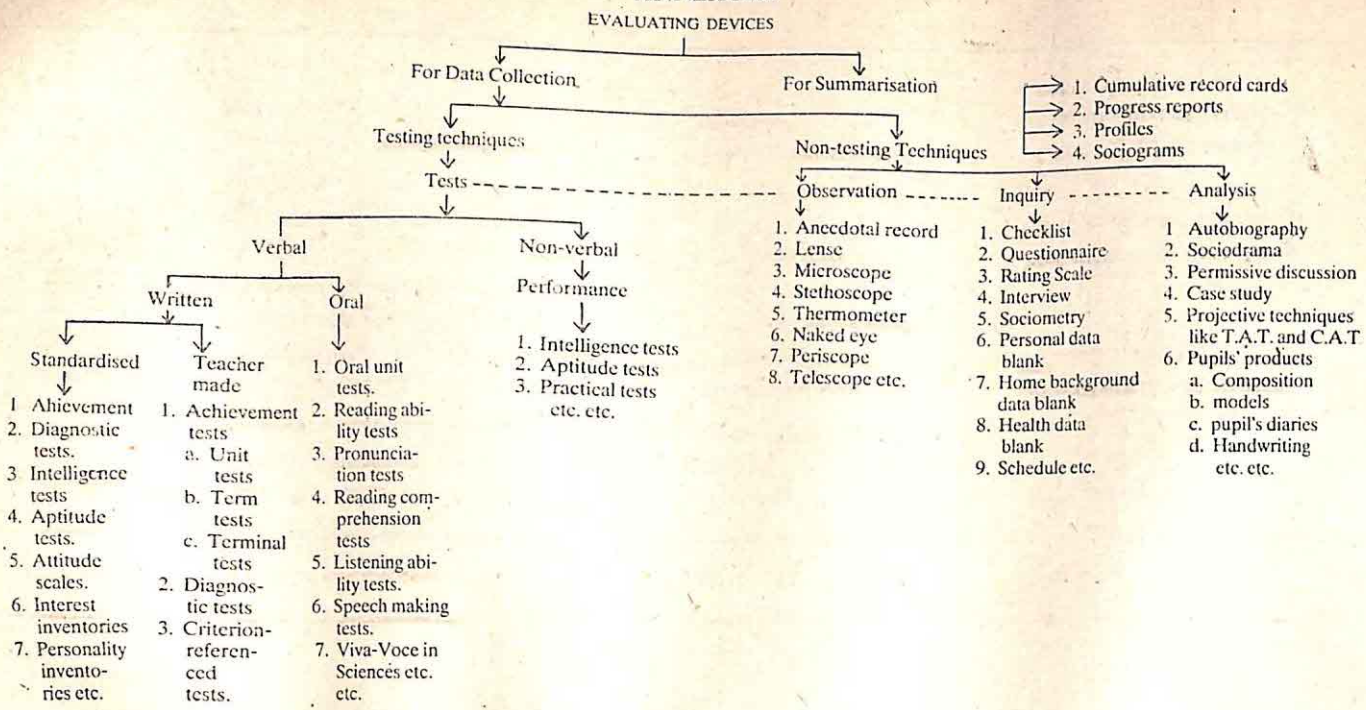
Analysis of documents, pupil's products, compositions, content, classroom questions etc. are all of great diagnostic value for improving instructional and learning strategies. This technique provides information about learning outcomes during the learning process in respect of mainly cognitive and psychomotor skills. It is quite objective in nature but not stable over time. It is inexpensive but preparation time is long, though crucial.

6. Relevance to Outcomes of Learning

To have a comparative view, we can summarise the data about all the four teachings in the tabulated form as under:

Technique	Types of evidences possible	Objectivity	Limitation
1. Testing	Student's achievement, Cognitive, Learning outcomes, Psychomotor outcomes, Attitudes, Interests.	Most objective and reliable More information available per unit time	Most expensive
2. Observation	Process of performance, Products of performance, Affective outcomes, Psychomotor outcomes, Typical behaviour, Social interaction.	Inexpensive. Usable in most of the situations. Provides direct observable evidence.	Subjectivity in assessment Time-consuming
3. Inquiry	Opinions, self-perceptions, Social perceptions, attitudes, subjective judgements.	Less expensive	Not stable over time. Lengthy. Preparation time but crucial
4. Analysis	Process goals, Psychomotor and cognitive skills.	Fairly expensive and objective	Least objective. Subject to bias and errors. Time-consuming.

ANNEXURE


EVALUATING DEVICES
For Data Collection
For Summarisation
1. Cumulative record cards
2. Progress reports
3. Profiles
4. Sociograms
Testing techniques
Non-testing Techniques
Tests
Observation
Inquiry
Analysis
Verbal
Non-verbal
Written
Oral
Performance
Standardised
Teacher made
1. Anecdotal record
2. Lense
3. Microscope
4. Stethoscope
5. Thermometer
6. Naked eye
7. Periscope
8. Telescope etc.
1. Checklist
2. Questionnaire
3. Rating Scale
4. Interview
5. Sociometry
6. Personal data blank
7. Home background data blank
8. Health data blank
9. Schedule etc.
1 Autobiography
2. Sociodrama
3. Permissive discussion
4. Case study
5. Projective techniques like T.A.T. and C.A.T
6. Pupils' products
a. Composition
b. models
c. pupil's diaries
d. Handwriting
etc. etc.
1. Intelligence tests
2. Aptitude tests
3. Practical tests
etc. etc.
1. Oral unit tests.
2. Reading ability tests
3. Pronunciation tests
4. Reading comprehension tests
5. Listening ability tests.
6. Speech making tests.
7. Viva-Voce in Sciences etc. etc.
1 Ahievement
2. Diagnostic tests.
3 Intelligence tests
4. Aptitude tests.
5. Attitude scales.
6. Interest inventories
7. Personality inventories etc.
1. Achievement tests
a. Unit tests
b. Term tests
c. Terminal tests
2. Diagnostic tests
3. Criterion-referenced tests.

CHAPTER VIII

GENESIS OF A QUESTION

1. Introduction

A question is a stimulus which evokes the intended response. According to the New Webster Dictionary (1), it is an inquiry, an investigation or an interrogative sentence soliciting an answer from an individual with some particular purpose. When applied to testing it may be defined as the smallest unit of a test. This word is sometimes used interchangeably with an item. Each statement of a true/false test or each blank in a completion test represents a question or an item. An exercise may consist of one or more questions. A question may be viewed as a single isolated question as well as a constituent unit of an exercise or a question paper. In either case the impact of a question would differ as also its characteristics. Thus the connotation of a question varies not only with reference to the context in which a particular question is used but also with reference to its existence as an independent unit or a part of a test. It is therefore, desired that as a result of going through this chapter the teacher will be able to

(a) understand the concept of a question,
(b) identify the bases of classifying questions into different categories,
(c) relate various mental processes involved in developing a paradigm of a questions,
(d) explain the genesis of a question,

(e) describe various characteristics of a good question, and
(f) apply relevant criteria to judge the quality of a question.

2. All Questions are Not Questions

The word question when used in the general sense does not in fact relate to an act of questioning. It may take different forms other than questioning. Let us observe the following examples:

(a) List three characteristics of living organisms.
(b) State three laws of motion.
(c) The capital of Mauritius is.....
(d) Justify the introduction of open book examinations at the Higher Secondary stage in your state.
(e) The Cabinet Minister for Human Resources Development is
  (A) V.P. Singh
  (B) Narasimha Rao
  (C) K.C. Pant
  (D) Krishna Sahi
(f) Why do you put a lot of salt to preserve pickles?

All these 'questions' in the real sense of the word are not questions. The first one requires enumerating; the second one requires only a statement; the third one demands completion of a sentence, the fourth concerns eliciting only positive arguments and the fifth one requires selection of the correct answer. It is only the last one which indeed appears to be a question. Thus if different types of so-called questions are categorised, we would find that all of them do not involve query and the process of questioning. However, the word 'question' in educational parlance has come to stay with the connotation of an item or a constituent unit of a test or an exercise demanding the examinees to *do* this and to *do* that and is not limited to an inquiry or questioning as in the case of the last question.

3. The Basic Source

The basic source of every question is the curriculum content or the syllabus prescribed. This does not, however, mean that one cannot frame a question independent of a syllabus. What I intend to convey is that every question has its root in some

subject-content. Therefore, when one thinks of a question in terms of its purpose one has to think whether that question is to be used for an instructional purpose or for an evaluation purpose. Accordingly, the question can be development-oriented or judgement-oriented. If it is development-oriented, then it can be used as a teaching device and if it is judgement-oriented it may be used as a testing device. However, the focus of both would be on measurement vis-a-vis attainment of an instructional objective.

4. The Context

The dynamics of every question varies from context to context. If the context is teaching, the purpose of a question could be developmental, diagnostic, review or improvement of the teaching-learning strategy. If the context is textbook, the purpose of a question may be enrichment, re inforcement, revision or furtherence of students' learning. If the context is examination, the purpose of a question may be to judge the attainment of an objective, to identify under-achievers, to classify the students for remedial action, to certify or to judge curriculum effectiveness.

5. Facets of a Question

There are umpteen ways of categorising the questions, depending upon the basis which one uses for classifying. In the context of testing, depending upon the purpose, a question can be used for improvement of students' learning, for selection or for certification of their achievement. In terms of function, we may use a question for grading, diagnosis or for reinforcement of learning. As a device, a question may be written, oral or practical. A question may be a constituent unit of the technique of testing, observation or inquiry. As a tool it may be in the form of a question, statement or an activity. It may be of the supply-type, selection-type or of the rearrangement-type depending upon the format of response. It may be categorised as free response, fixed response or restricted response depending upon the scope of expected responses. Its content of testing may be a fact, a concept or a process. It may be based on a knowledge, an understanding or an application objective. It

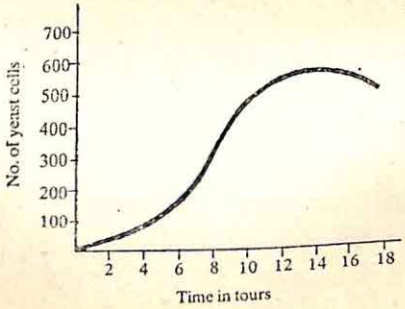
may represent molecular, cellular or ecological levels (in case of a bio-question). It may be set at an easy, average or a difficult level.

Thus in a testing situation (context) a question can be set for grading (function) students to certify (purpose) their achievement with the help of written words (mode) using a diagram (medium) by means of the supply-type (format) demanding restricted response (scope) of short answer-(form) type based on a concept (content element) to judge students' ability to interpret (objective) and can be attempted by 50% to 60% of students (average difficulty). Most of the hidden facets of a question as indicated in the above statement can be

Facets of a Question

Context	Instruction	Testing	Textbook exercise
Function	Diagnosis	grading	Reinforcement
Purpose	Improvement	Certification	Selection
Mode	Oral	Written	Performance
Medium	Word	Diagram	Activity
Format	Selection type	Supply type	Rearrangement type
Scope	Extended response	Restricted response	Fixed response
Form	Long answer	Short answer	Objective type
Content element	Fact	Concept	Process
Objective	Knowledge	Understanding	Application
Difficulty level	Easy	Average	Difficult

Example : What does this graph indicate regarding growth of yeast cells



depicted in the form of a scheme given on the previous page.

6. The Paradigm of a Question

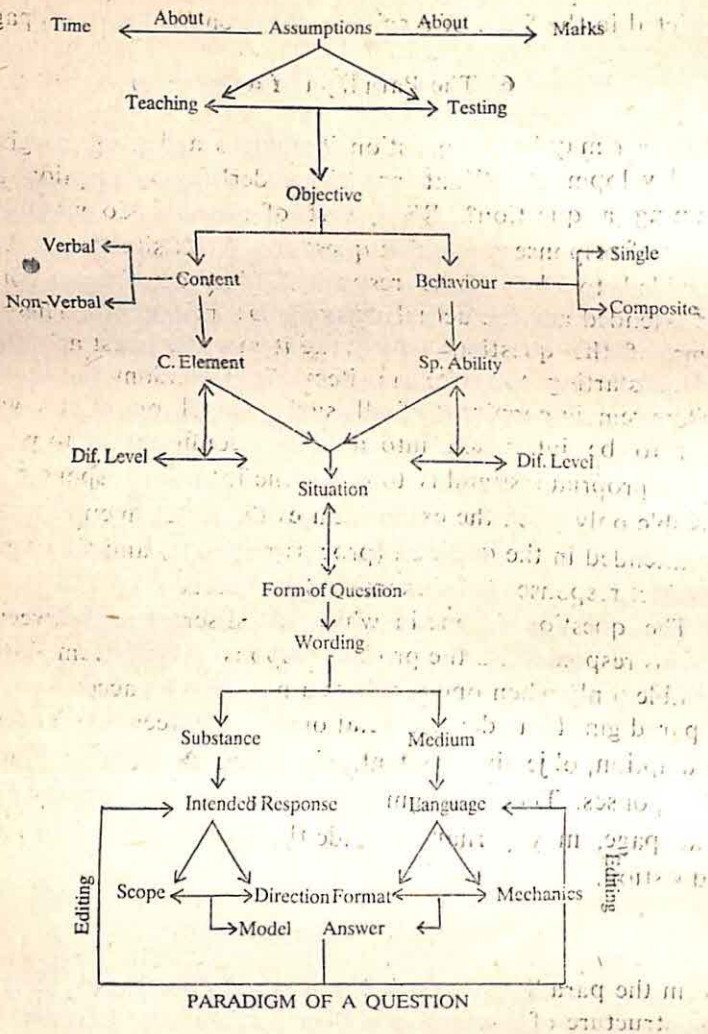
Whatever may be the question it requires a lot of insight for its development. What are the underlying assumptions while framing a question? What sort of answers do we except in terms of responses to such a question? What situation is to be provided to elicit those responses? How can the gap between the intended and the actual responses be minimised. These are some of the questions which the item writer must ask himself before starting to write an item. It is seldom that the item writers remain cognizant of all such mental operations which need to be integrated into a desired configuration to provide the appropriate stimulus to evoke the intended responses. It is possible only when the examinee uses the same mental process as intended in the question (process response) and the response (product response of the examinee) is correct.

The question is one in which the discrepancy between the process response and the product response is minimum. This is possible only when one develops a new item in accordance with a paradigm that depicts relationship between the underlying assumption, objective, content, situation, wording and nature of responses. The paradigm given in a diagram appearing on next page, may perhaps provide the needed genesis for writing a question.

7. Genesis of a Question

From the paradigm given below it is not difficult to infer that the structure of a question is closely related to its functionality. What usually happens is that the phenotype (outward appearance) of a question does not always reflect its genotype (genetic material or constitutional make-up).

Thus a question has two faces; one which is manifested in the form of observed format i.e. wording of stem and alternatives, say in a multiple choice question (phenotype). The other face is represented by its built-in characteristics (genotype) of the question forming its substantive structure. This structure comprises a set of characteristics which are regarded



PARADIGM OF A QUESTION

as symptomatic of its potentiality to evoke the desired responses in a testing situation. When the phenotype (manifested responses) is the same as embedded in the genotype of the question it is a case of complete congruence between the intended and observed responses. However, phenotypic (observed) responses of a question may be different inspite of the same genotype. This is a case of discrepancy between the intended and the observed responses. Likewise the genotype of two questions may be different but the phenotypes (observed responses) may

be the same. This incogruency may be due to poor wording of a question or misconception of the objective being tested. The best question is one whose phenotype is congruent with its genotype. The following table would indicate the various genotypes of a typical question reflecting three of its major dimensions, namely, the objective, content element and form of question.

K=Knowledge	F A E	C A E	P A E
U=Understanding	F A S	C A S	P A S
A=Application	F A O	C A O	P A O
F=Fact	F U E	C U E	P U E
C=Concept	F U S	C U S	P U S
P=Principle	F U O	C U O	P U O
O=Objective type	F K E	C K E	P K E
S=Short answer	F K S	C K S	P K S
E=Essay	F K O	C K O	P K O

Genotypic table of a question (cognitive domain)

The three basic characteristics which forms the genotype of a question in a cognitive domain are the 'content', 'objective', and the 'form' of question. The content may be a fact (F), a concept (C) or a principle (P). The objective may be knowledge, understanding or application. The form of a question may be the objective type (O), the shorter answer type (S) or the essay type (E). Thus we have ($3 \times 3 \times 3$) genotypes of a typical question leading to 27 types of questions which can be framed. This includes 9 varieties of knowledge questions, 9 of understanding and 9 of application as shown in the above table.

8. Characteristics of a Good Question

For framing a good question, therefore, we have to choose the relevant content, objective and form of question. This does not, however, ensure that cognizance of these three would serve the purpose for which it is developed. It has to be couched in such a language that it provides the needed direction to the students

to evoke the intended responses. The first criterion of a good question therefore, is its relevance to the intended genotype. Secondly, it must evoke intended responses in terms of its genotype. Thirdly, it must provide responses which are acceptable at a particular level in terms of its criterion. Fourthly, it must ensure the maximum congruence between the process responses (intended) and the product response (observed). Fifthly, in terms of its efficiency, it should yield appropriate facility and discrimination indexes. Lastly, it must be economical in terms of time requirement for maximum coverage. Whatever may be the criteria of a good question, the best question is one which evokes those very mental processes to attempt is correctly, which are intended by the authors of the question.

9. Dynamics of a Question

A question plays a significant role in the process of teaching and learning by providing the basis for interaction among its users. It can engender interaction among students, teachers, peers, parents, paper-setters, examiners, examinees and administrators. A question can be used by the students for self-learning revision of a lesson or for self-evaluation. A teacher can use it for development of a lesson, for reviewing a lesson and for diagnosing inadequacies in learning. A paper-setter can use a question for setting a standard, classifying pupils and certifying pupils' achievement. An examiner may look to acceptable responses, uniform scoring and objectivity in marking. An examinee tries to get insight into the intended response, understand the intent of the author and guess the nature of the intended response. The peers try to discuss the expected answers, compare their responses with the intended ones, and use it for self-evaluation. Parents judge their wards' performance to identify gaps in their' learning and coach their wards in terms of expected questions. Administrators judge performance standards of pupils, quality of examination and effectiveness of the instructional process. (2)

Thus the interaction indicated above highlights the importance of a question which forms the basis of interaction among

its clientele. Therefore, when a question is used in different situations it provides interaction of various types. This is however, possible only when its user employs it at the proper time in an appropriate manner for a predetermined purpose in the given context.

10. Judging a Question

All methods of judging a question are subjective. Moreover a question when considered as an independent entity and when it is a constituent unit of a test, manifests its behaviour differently. A question may not turn out to be a good question inspite of the best reflective judgement made by experienced teachers and subject specialists. This is because every individual has a different background and technical know-how of the content element and the objective being tested. The same question is characterised as knowledge based by some and as application based by others. This again depends on the assumptions made by the critic about the quality of instruction provided to the students. When a question is tried out to get the empirical validity, the same problem arises because it gives different results with different groups of students taught by different teachers. A question in terms of facility index and discrimination index may be very good as an individual question used with a particular class. But the same question when judged in terms of a different class, may lose its potential and rated as more difficult or more easy, having lesser or more discrimination index, the question remaining the same.

In a norm-referenced test, a question with a low discrimination index of, say, 0 1 may be considered unsuitable because it gives very little variance between the low and the high achievers. But when the same question providing the same index of discrimination is visualised as a part of a criterion-referenced test, it may be considered suitable because it shows low variance, which indeed is the focus in such tests for identifying the masters and non-masters. Similarly, all good questions when put together in a question paper do not necessarily make it a good one. In fact, it may result in a bad paper: for example, when all good application questions are included in a

question paper meant for a heterogeneous group. An easy question paper may lower its validity for one group of students but may prove difficult for a different group and vice-versa. An application question proves to be a simple knowledge question for students taught with emphasis on higher level objectives.

Thus a good or a bad question is good or bad in its own niche or context. The quality of a question, therefore, can be judged only in terms of a particular group of students and the quality of instruction they received. One cannot judge the quality of a question only through test results. The best way, therefore, is that the classroom teachers should construct quality items, with the help of fellow teachers teaching the same subject in the same institution, followed by the use of those questions in classroom examinations and improving them further through analysis of responses of examinees. Once the teachers are well trained in the technique of item-writing, long-term and short-term targets can be fixed for development of item pools which can be used later for teaching and testing.

11. To Sum Up

A question is a useful but dangerous tool in the hands of incompetent teachers and evaluators. For its effective use in teaching, learning and evaluation, one has to understand fully its dynamics. A bad question is a reflection on the author's limitation of technical competence and mastery of the structure of the discipline. Unless the item writer is able to integrate content elements and specific abilities properly, it will not be possible for him to produce a good question. Using a badly framed question is professionally unethical as it creates mistrust and wrong conclusions about instructional efficiency and the pupils' learning. Framing a good question is not everybody's cup of tea and passing judgement on the quality of a question without knowing the context, focus and purpose is unethical. I have tried to raise a number of questions about a question in this paper.

CHAPTER IX

CONSTRUCTED QUESTIONS—(SUPPLY TYPE)

A. LONG-ANSWER QUESTIONS

After going through this chapter it is expected that the reader will be able to

(a) know the difference between objective based and objective type questions,
(b) classify the given questions into Long answer, Short answer and objective type using the prescribed criteria.
(c) give examples of extended response and restricted response type of questions testing different abilities.
(d) improve the defective questions of each type by using the principles of framing questions,
(e) analyse the factors affecting subjectivity in making long answer questions and the associated steps for improving their scoring objectivity.
(f) construct short answer and very short-answer questions using relevant criteria to test different abilities.
(g) detect errors in the given questions of short answer and very short-answer questions and improve them using the relevant principles of construction.

1. Objective-based Versus Objective Type Questions

In the previous pages we have described the various tools and techniques that can be used to obtain evidences about pupils'

growth in terms of expected outcomes of learning. By and large teachers are more concerned with the collection of evidences in terms of instructional objectives towards which all-out efforts are made in the classroom using varied types of teaching-learning strategies. How many teachers take cognizance of the objectives or the intended learning outcomes while writing a question, is anybody's guess. It is, therefore, essential that before framing a question, the need for identifying the intended objectives cannot be over-emphasised. Indeed, every question is based on one or the other objective but the fact remains that the item writer is seldom conscious of the objective he is going to test for. A question may require the students to write a long or short answer or just select an answer out of the 4 or 5 designated responses, as in the case of an objective type question. An objective type question refers to the form of the question whereas an objective-based question refers to the base of the item i.e. the instructional objective on which it is based. Thus an objective-based question may be of the long answer, short answer or the objective type variety. Every question of the objective type is, therefore, objective-based but every objective-based question need not be of only the objective type.

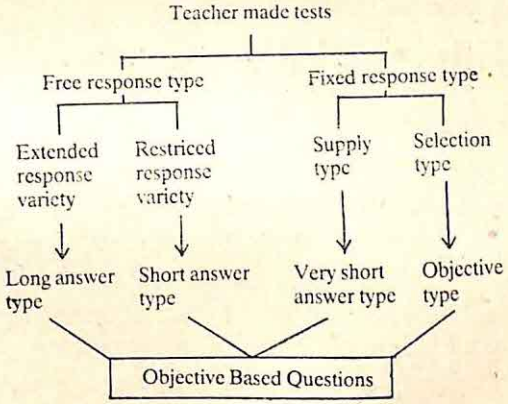
1.1. *Classification of Question*

Depending upon the nature of expected responses and mode of attempting a question, we can classify all questions into two categories viz, supply type and selection type, the former requiring students to supply the answers while the latter demanding only the selection of correct responses out of the given ones. Since the nature of responses in the supply-type questions would vary, ranging from the very extended response variety to that of the very restricted form requiring students to supply just one word. We may classify them on a continuum of response length. The following list of questions would further illustrate this viewpoint.

Questions:

1. Write an essay on the uses of plants.
2. Write an essay on the uses of medicinal plants.
3. Describe four uses of medicinal plants.
4. Describe two uses of medicinal plants with examples.
5. List four main uses of medicinal plants.
6. List four plants of medicinal use.
7. Name any two medicinal plants.
8. Name one medicinal plant.
9. Which medicinal plant gives quinine?
10. Which of the two plants, Cinchona or Citrus gives quinine?
11. Does the Cinchona plant give quinine?
12. Which of the following plants yields quinine?
    A. Aspergilus
    B. Cinchona
    C. Neem
    D. Citrus

From the above examples, it is clear that on the basis of restriction imposed on the degree of freedom of response, one can classify various types of questions as depicted below:



All these forms of questions are based on the nature of responses and the mode of answering the questions. Each of these types of questions is discussed hereinafter.

2. Essay-type Questions

The most commonly used tools of evaluation are the essay-type questions, although the other two types viz., short answer and objective type are gaining prominence day by day due to their inherent advantage of being scored more objectively and interpreted more easily. Despite the wider applicability of the short answer and objective type questions, there are certain outcomes of learning e.g. to organise, summarise, integrate ideas and express in one's own way which cannot be satisfactorily measured through objective type tests. The importance of essay tests lies in the measurement of such instructional outcomes.

Essay tests have become the focus of criticism for those who are interested in the science of measurement as in achievement. Due to this, essay tests have been left undeveloped inspite of the fact that they are still the most widely used form of test by teachers. They have some unique, potential educative values. Therefore, the need for developing these tests and undertaking research in this field cannot be underrated.

2.1. *Two Varieties of Essay-type Questions*

Essay tests may be of the extended response type or the restricted response type depending upon the limits of the expected response. An essay question, unlike an objective question, is not an all-or-none affair but rather a matter of degree. On the one extreme, an essay question may give full freedom to the students to write any number of pages. On the other hand, a restriction may be imposed on the length of the response by limiting the scope of the expected answer. This restriction may be imposed to such an extent that the required response may vary from one word, a phrase or at the most a sentence, a few lines or even a few paras. This limit may be imposed by restricting the content and the length of students'

response in the statement of a question. Restricted response type items are quite useful for learning outcomes which require interpretation, application of data or outcomes which are specific and clearly defined in nature. Such type of questions help to reduce subjectivity in marking, for which essay tests are infamous.

In extended response type questions, full freedom is given to the student to exercise his competence and demonstrate the best he possesses, of course pertaining to the area of the subject. He is free to select, organise, integrate, evaluate and express in any way he likes or deems appropriate. Such questions although useful for measuring global type of abilities, are not suitable for measuring specific learning outcomes, besides being difficult to grade.

2.2. *Classification of Essay Questions*

Monroe and Carter (1923) divided essay questions into 21 types (3): They include questions on (1) Selective recall, (2) Evaluative recall, (3) Comparison of two things—single basis, (4) Comparison of two things, general, (5) Decision—for or against, (6) Cause or effect, (7) Explanation of the use or the exact meaning of the same word, phrase or statement in a passage, (8) Summary of some unit of the text or of some article read, (9) Analysis, (10) Statement of relationship, (11) Illustration or examples (the pupils') of principles in science, construction in language etc., (12) Classification, (13) Application of rules, laws or principles to new situations, (14) Discussion, (15) Statement of an author's purpose in his selection or organisation of material, (16) Criticism, (17) Outline, (18) Reorganisation of facts, (19) Formulation of new questions, problem and question raised, (20) New procedures, (21) Interartial thinking.

Weidman (1923, 1941) distinguished 11 definable types of tests arranged in series from the simple to the complex with directional words as under (4):

(1) what, when, who, which, where, (2) list, (3) outline, (4) describe, (5) contrast, (6) compare, (7) explain, (8) discuss, (9) develop, (10) summarise, and (11) evaluate.

2.3. *Requisites of Essay-type Questions*

(a) Essay questions should be set to test only those instructional objectives which are not amenable to testing by other forms.

(b) Each question should be based to test specific mental processes, learning outcomes implied by the objective in view.

(c) Phrase questions in such a way that their meaning and intent are clear to the examinees.

(d) Structure questions in such a way that the scope of the expected answer is clear.

(e) Directional words like, 'what do you know of', 'give an account of', 'write short notes on', may be avoided to avoid vagueness of answers and consequent subjectivity of scoring.

(f) Maturation level of the examinees must be taken into consideration while constructing an essay question. The length and nature of answer will differ from class to class. For example, questions requiring discussion, interpretation, summarisation and evaluation may be asked in higher classes whereas questions like listing, describing, selecting etc. may be considered for lower classes.

(g) Marks should be clearly allocated part-wise, whenever there is more than one part in the same essay-type question.

The realibility of essay examinations can further be improved if the students trained properly through the use of the important variety of such questions in day-to-day testing programmes in home examinations. They need to be familiarised with the method of attempting such questions in accordance with the connotation of the various words, especially the directional words used to circumscribe the nature and scope of the answer expected. This will ensure, to a great extent, the consistency in each student's understanding of what he is required to write in response to a particular question and the way the teacher is going to grade it.

2.4. *Improving Essay Questions*

Improvement of essay questions may be made keeping in view the factors that affect the measurement value of this instrument both at the level of framing the questions and at the level of scoring the scripts. This can be taken care of by taking some precautions of a general nature regarding planning of questions and their grouping in the form of a test and certain specific criteria for structuring the questions for ensuring better reliability in marketing. (4)

2.4.1. *General Hints*

(a) Give as much time to writing a good essay question as it would require to score it. This would help to remove many of the angularities of questions that create trouble in marking.

(b) Avoid choice of questions to discourage selective study, ensuring better content coverage and making comparison of students' performance easier.

(c) Structure each question keeping in view its time requirement.

(d) Keeping in view the range of difficulty, value as wide as possible to cater to the students of different abilities.

(e) Eliminate as far as possible all irrelevant factors like spellings, handwriting etc., that affect scoring. Make specific provision for these if they are to be specially emphasised or considered at all.

(f) Write explicitly the general directions for questions to provide a common base for all pupils to attempt questions on similar lines towards the same goals.

(g) Try to increase the number of questions by including more questions of the restricted response type for securing more content coverage.

(h) Use essay questions where you must and avoid where you can. They may be utilised only for those objectives and content areas which are amenable to testing through such questions.

2.4.2. *Specific Hints*

(a) Base your questions on the selected instructional objectives.

(b) Make use of the specific content element appropriate to the purpose of the question.

(c) Structure question so as to delimit the scope of the expected answer to pinpoint the area of response.

(d) Use simple, precise and unambiguous language to avoid semantic difficulties.

(e) Indicate marks clearly, part-wise, if the question consists of more than one part. Moreover, these marks should be in whole numbers and divisible according to the number of credit points. This ensures more objectivity in scoring.

(f) Set tasks in the question which require the students to demonstrate their command of essential knowledge.

(g) Use familiar directional words and those which elicit the required responses e.g. compare, give reasons, etc. Directional words like 'Write short notes on', 'What do you know of', 'Give an account of' etc., should be avoided as they create difficulty in understanding the nature and scope of the expected answer. Directional words like 'discuss', 'comment', 'justify' should be used less frequently and only after having trained the students in the use of such words in relation to the type of responses expected in such cases.

(h) Test your question by writing a model answer. Writing the expected answers helps to improve the question by serving as a check on the language used, scope of the question, allotment of marks and the objective tested. This leads to the remodelling of the question and/or the answer to fit each other.

Example

Let us take an example to indicate the various steps involved in writing a good essay question. Before testing students on the unit 'Organic Evolution', Miss Biology gave some thought to the type of essay type question she could frame to test student's achievement on the unit. She framed the following question:

(i) Write an essay on organic evolution.
It looked quite acceptable at first but at a second look she felt that it would be too lengthy a question where students can write anything about the concept, evidences, theories, mechanism of inheritance etc. What she wanted was to test students' ability to analyse various theories of organic evolution. So she reworded the question as under:

(ii) "Write an account of various theories of organic evolution."
On further reflection, she realised that in this revised question there is over-emphasis on the recall of factual information whereas in the blueprint of the unit test she wanted a question involving analysis and judgement-making abilities. She tried to improve it further and reformulated the question as under:

(iii) "Describe the various principles that underlay various theories of organic evolution."
This question was also rejected by Miss Biology as she felt that it did not really invoke students' analytical skills. Therefore, the question was reworked as follows:

(iv) "Which theory of organic evolution appeals to you the most and why?
This question was considered nearer to the intended objective as viewed by the framer. However, on close scrutiny it was thought that what Miss Biology wanted was to judge students' ability to argue for and against various theories with regard to the adequacies or inadequacies in explaining the mechanism of organic evolution. She therefore, reframed the question as follows:

(v) "Which theory in your opinion explains better the mechanism of organic evolution and why?"
Even this question did not satisfy her completely when viewed in the light of its scope, which remained limited to testing of only one theory rather than providing for evaluation of each theory to pass a judgement. Thus the question was reworked and reframed as under:

(vi) "Discuss the various theories of organic evolution from the point of view of the mechanism of inheritence."

Miss Biology reached the point where she was happy that her question now satisfied the major criteria of a good essay question. Nevertheless, what still remained is the working out of the outline answer and value points which would form the basis for the marking scheme. Moreover, she had also to inform students whether, handwriting, spelling and grammatical mistakes would be considered while grading the answers.

All the above mentioned steps involved, relate to the cognizance of instructional objectives, the scope of expected response, the use of the appropriate directional word, precise language, intended content-coverage and ensuring maximum possibility of scoring objectivity.

2.5. *Grading Essay Questions*

The effectiveness of essay questions depends largely on how well they can be graded. An ill-conceived and poorly-developed essay question cannot be salvaged by even the most carefully devised method of grading. Improper grading however can spoil the well-constructed essay question. Nevertheless, unreliable grading is one of the most valid criticisms levelled against essay questions and, therefore, we need not over-emphasise the scoring objectivity of the essay tests. To grade responses on essay questions we can try to:

(a) identify appropriate methods to minimise biases,
(b) attend to significant and relevant aspects of the answer,
(c) apply uniform standards of scoring to all scripts,
(d) prevent personal idiosyncrasies that affect grading.

We should aim at optimum reliability only and try to reduce intra-examiner and inter-examiner variance. Scoring unformity and scoring objectivity should be ensured as far as possible by taking necessary steps at the time of setting the question and while developing the marking scheme. Two methods of grading can be used (5):

2.5.1. *Grading by analytical method*
This method is also called the point score method in which a model answer is broken into specific value points each of which is given proportional weightage in allotment of marks.

2.5.2. *Global scoring method*
It is sometimes also called the rating or holistic method. In this method the ideal answer is not split into component parts or specific points, but it simply serves as a standard. The responses of students are graded on the quality continuum which has a number of standards or anchor points. The rater reads the response, forms a general impression and assigns a rating, using some standard on the quality continuum. A response can be graded as good, average or poor. To be more discriminating, it may be graded on a 5-point scale like excellent, good, average, satisfactory and poor. Such scale values can be established by preparing a variety of answers corresponding to the various scale points. Alternatively, the scorer can select papers from the already written actual responses to establish the various anchor points. In a five-point scale, the responses can be categorised in terms of:

1. superior quality,
2. above average quality,
3. average quality,
4. below average quality, and
5. inferior quality.

The reader after a rapid reading, can assign scripts into 5 piles. To ensure more objectivity each response should be read and classified twice, preferrably by a different rater who assigns the grade independently. When a large number of essays are to be graded, the global method is more useful especially for the teachers who have only 30-40 scripts for grading. However, if the number is extremely small, it will be difficult to establish anchor points. In case of a homogeneous group of students, we can make preliminary reading and assign the paper into one of the piles. Each question can then be re-read one or two times to reclassify the scripts which are

misclassified.

A student's score depends on the number of value points contained in his answer. Where expression is involved, related value points like those of logical organisation, effectiveness of expression, arguments to support or reject an issue etc. are specified and assigned points of values. Thus a sort of listing of value points in the form of a checklist is done for the use of examiners. Depending upon the complexity, significance and length of the expected responses, different values for different points can be given e.g. a credit point may carry one mark while another can carry two or three marks. As the scorer reads the responses, he goes on giving points to those credit points which are contained in the answer.

This method minimises the influence of extraneous factors. Despite a detailed scoring guide, it is not possible to control human variability because of the fallibility of the human teacher i.e. the scorer. However, this method yields very reliable scores when used by a conscientious scorer. The outline answer and detailed marking scheme, when prepared alongwith the framing of questions, helps to remove a number of incongruencies such as the difficulty level, unrealistic time, wording etc. The five sub-divisions of a model answer make it easier to assign grades to students which are considered more trustworthy. It is of course a laborious, time-consuming method and may lead to identification of aspects or elements which may be superficial. This method is more suitable to apply in case of the restricted response variety of essay questions which demand more specific responses. The extended response variety of essay questions can be scored by this method more reliably for the same reason.

3. Improving Reliability of Scoring

The following precautions may be taken to control the reliability:

(a) Use precise, unambiguous and understandable language to enable the examinees to make the same meaning out of the question so as to focus their attempt at the same time.

(b) Delimit the scope of expected response by structuring the question in such a way that it pinpoints the area of response desired.

(c) Use appropriate directional words to avoid semantic difficulties. A particular directional word should be carefully selected in conformity with the abilities to be tested. This helps to improve validity by directing the purpose of the question, thereby ensuring more reliability.

(d) Allocate marks part-wise where the question reflects different aspects to be credited separately. It helps the examinees to lay proportional emphasis on each part and facilities scoring on the part of the examiner.

(e) Make use of proper directions in the question paper to enable the students to know whether organisation, handwriting, brevity etc. in the answer would be given some credit or not. This will also reduce unreliability.

4. Improving Scoring Objectivity

Following suggestions may be considered useful to reduce subjectivity in scoring essay type questions (6):

(a) Write an ideal answer to the question in advance. This should include major points to be evaluated. These are the value points expected from the students. In extended response items, the main points may be outlined. This provides a common base for grading the scripts, thereby reducing subjectivity in scoring. Model answers prepared for each question at the time of construction of items helps to phrase the question more clearly to convey the type of answer expected.

(b) Evaluate answer scripts as anonymously as possible. The impression a student makes a boy causes persistence effect while marking the scripts. To avoid this type of bias either some fictitious roll number may be allotted or students may be asked to put their names at the end of the answer script so as to conceal their identity till the whole answer script is examined.

(c) Use the analytico-synthetic method for grading. This in-

volves scoring by the point method or the rating method. In case of restricted response questions, the point method involving analysis of the various value points may be used by comparing the responses with the ideal answer prepared in advance. For global-marking, the rating method may be used by placing the answer scripts in different piles, say five, according to the quality of responses. This method suits evaluating questions of the extended response type where gross judgements relating to organisation of ideas, their relevance etc., are to be made. Whenever the latter method is used, it is profitable to rate each characteristic separately—organisation of ideas, relevance etc. In the common 100 point scale is too precise to be used and that is why it may be desirable to use a 5 point or 9 point scale. Moreover, instead of a point score we may use the range score or the band score based on the score and the standard error of measurements. For example for a score of 50, if the standard error of measurement is 2, and we take deviations upto three sigmas, we have a band score—44-56.

(d) Use question-wise scoring rather than script-wise scoring. When different scripts with different quality levels of responses are piled after going through them hurriedly, then scoring of each script may be undertaken. It is better to evaluate one question through all the scripts before passing on to the next question. This helps in using the same basis for judgement for all questions through constant comparison of the same type of responses. This reduces the shifting of standards from one paper to another paper. When this method is used, there is less likelihood of the "persistence effect", which is always formed when we examine script by script.

(e) Score the question for each aspect separately according to the part-wise allocation of marks. This will tend to enhance scoring objectivity.

(f) Mark some scripts for the second time for comparison with the original score to confirm reliability of marking. A few scripts in the beginning may be left unmarked and a separate record of the score may be kept for this purpose.

(g) Reshuffling of answer scripts after marking each question helps to avoid assessment of the same scripts in the same order thereby reducing the carry-over effect.

(h) Double grading may be used when the test is to be used for selection or awarding scholarship etc. Otherwise, only a few scripts may be rated independently by two examiners for better comparison and to confirm reliability of scoring.

(i) Total the marks of the scripts at the end after marking all the scripts. This will avoid carry over of any impression that is formed about the quality of the script when marks are totalled.

(j) Avoid creeping in extraneous factors like spellings, handwriting, punctuation, sentence structure, neatness etc., during marking of the scripts. Such outcomes if stressed for their own sake, should be evaluated separately and specific marks may be kept apart for independent assessment. This would tend to avoid contamination of of test scores which indicate achievement of other specific outcomes of learning.

(k) Examiners may meet together after marking some scripts to discuss the marking scheme and comparing marking. This results in reducing inter-examiner variability.

(l) Although it may be a little costly to provide separate answer sheets for each question, there is no denying the fact that it can bring improvement in scoring objectivity by allocating scripts question-wise to the examiners. As each examiner would be marking only one question scoring, objectivity is sure to improve.

Essay examinations inspite of some inherent weaknesses have come to stay and are the major tools of evaluation used in internal as well as external examinations. Their significance for measurement of specific learning outcomes, which are not measurable by other types of examinations cannot be overlooked. What is needed is further development and research in this field to improve reliability of responses and accuracy of scoring.

(m) Be consistent in grading the scripts. There are other problems that need to be researched upon. In fact problems relating to reliability of essay questions are well known.

Studies conducted by Harper in India are worth quoting especially the two famous ones, the 'Ninety marking Ten' and "Four thousand re-examined". (7)

His findings are eye-openers both to the examining agencies and the policy makers in the field of school education. Those who are interested in knowing details of these studies may refer to the volume on 'Researches on Examinations' by A.E. Harper published by NCERT (1975).

Despite various limitations, the essay type questions have come to stay and their role in measuring certain outcomes of learning cannot be underestimated. What is needed is to improve the construction of these tests and scoring procedures with a detailed marking scheme. Well-designed essay questions with a well structured scoring procedure could enable the teachers to grade the students' responses more objectively, besides improving their validity and reliability.

B. SHORT-ANSWER QUESTIONS

1. Concept

Short-answer questions are open-ended questions with highly restricted responses. These are free-response type questions in which the response may vary from one word to a few lines depending upon the degree to which restriction is imposed on the expected answer. Such questions are ranked in-between the essay and the objective type question as regards the objectivity of scoring. But this does not mean that they are exactly midway between the two extremes. In fact, these questions can be placed on a continuum ranging from the highly objective scorability to the highly subjective scoring of responses. Thus they merge with either type at the ends based on the limits imposed on the freedom of response. Such tests become practically fixed-response type like any other objective type items when the specificity of response is restricted to one word. For example, when we put the question, 'Who invented the telephone?', it is almost an objective type question. On the other hand, a ques-

tion like, "Why do you feel more hot in summer on a day when the relative humidity is comparatively high?", is supposed to be a question of the short-answer variety but it is very close to an essay type question because its explanation may demand an answer whose limits may not be determined from the language of the question.

Although, such a short-answer question variety as a separate class or form of question is difficult to justify, yet for the convenience of scoring facility, a better content coverage and other practical considerations, such questions can be categorised separately. Nevertheless we may fix some arbitrary limits of such questions for practical purposes.

2. Scope

In order that such questions of the short-answer variety may yield better reliability, the following limits may be imposed on the freedom of response with reference to the allocation of marks and time requirements. (8)

(i) *Number of words*
Short-answer questions may be set to get responses which may vary from 30 to 50 words or so.

(ii) *Number of lines*
Although it is preferable to limit the response in terms of words rather than lines which may vary from individual to individual, yet we may restrict it from one to five lines presuming that each line contains 6 to 8 words.

(iii) *Number of credit points*
While framing short answer questions it is desirable that expected response may be limited to two, three or four value points. In case the examiner is interested in testing more credit points on the same content he should either set an essay type question or frame two or more short-answer questions on that content.

(iv) *Allocation of marks*
As the responses sought are restricted to a few specific value

points and the content area tested is also comparatively small, short-answer answers carry few marks than essay type questions. Usually one to three marks should suffice for a short-answer question.

(v) *Time requirement*

Depending upon the length of the answer and the credit points, the time required for each question may be estimated. Normally, a short-answer question should not require much time to answer. As the answers vary from one word to about fifty we may allot 1 to 5 minutes as the time limit for short-answer questions. All the limits imposed are arbitrary and based more on the practical experience rather than on any sound academic rationale.

3. Two Varieties of Short-Answer Questions

Sometimes short-answer questions are classified as selection type and supply type. The former type is named the objective type while the latter type may be further divided into two categories. Depending on the nature of expected response we may call it 'Very short-answer questions'. But when the response extends to about 30 to 50 words or so involving more value points than one of these questions, it may simply be called a 'Short-answer question'. Questions provided on the next page would illustrate the difference in the variety of short-answer questions indicating the nature and scope of expected responses.

The list of questions on the following page is not exhaustive but only illustrative to indicate the extent of the freedom of response allowed to the examinees. However, in all cases the restriction on length of response and specific credit points is imposed. In the last three questions (6, 7, 8) the emphasis is one value point. Question No. 5 could also be categorised as the very short-answer (VSA) type but because of two value points I have preferred to classify it as short-answer. For the sake of clarity and variety in construction, short-answer

Questions

1. Describe in about 50 words the functions of various floral, whorls in a typical flower.
2. Mention the names of different floral whorls of a typical flower giving the functions of each.
3. List one major function of each of the four floral whorls.
4. Give the technical terms for each of the four whorls of a typical flower.
5. Write two technical terms representing the essential parts of a typical flower.
6. The main functions of sepals is to.....(fill in the blanks).
7. Name the whorl representing the male part of a typical flower.
8. Which of the two, calyx or corolla, is more helpful to a flower in pollination?

questions can be further sub-classified as under:

(a) Open ended—	Restriction on response length but not defined	—Credit points not stipulated (Q. 1, 2)
(b) Open ended—	Response length defined	—more than one credit point stipulated (Q. 3, 4 & 5)
(c) Open ended—	Response in one word, one phrase or one sentence	—Single credit point (Q. 6 & 7)
(d) Open ended—	Response choice is provided in question	—One credit point stipulated (Q. 8)

Thus all these types of short-answer questions can be arranged on a continuum of freedom of response which admit varying degrees of scoring objectivity. On the one extreme they admit subjectivity both in content and organisation of responses (Q. 1) while on the other it ensures 100% objectivity in scoring by mere selection of the correct response from the two designated ones (Q. 8). Depending upon the nature and content of a unit and the extent of scoring objectivity desired, one can make a choice to include them in a test.

4. Hints for Construction of Short-Answer Questions

Short-answer questions, if constructed properly, can be taken advantage of in overcoming many of the limitations of the essay and objective-type questions. The following points if kept in view while constructing short-anwer questions would surely go a long way to improve their quality: (9)

(a) Use these questions for testing only those objectives which are amenable to this type. These tests, for example, may be used to test functional information, understanding and application of knowledge to a great extent. Even certain drawing skills representing specific areas or points can be tested by using short-answer questions. However, such questions cannot be used profitably to test mental processes like ability to organize, marshall facts, develop a prolonged and integrated argument for which essay type questions are most suited. Almost all specific objectives can be tested by the use of these questions as shown in the annexure given at the end.

(b) Frame questions in such a way that maximum coverage of content is ensured. This is very essential to utilise each mark allotted to such questions by getting the maximum benefit of adequate coverage of content. For example a question carrying one mark may be put in either of the following forms:

(i) Mention one property of oxygen.
(ii) Mention one chemical property of oxygen.
(iii) Mention one chemical property of oxygen which is different from that of carbon dioxide.
(iv) Mention one property of oxygen which is different from that of carbon dioxide but resembles that of hydrogen.

The first question tests almost nothing as one can write simply colourless/odourless/tasteless. In the second question content coverage is more than in the first because the examinee must know the difference between chemical and physical change before answering it.

In the third question there is still more content coverage involving two gases. In the fourth question the content coverage is still more to cover the properties of the three gases. It is this question which ensures maximum coverage, utilising just one mark.

(c) Try to cover 2, 3 or 4 credit points instead of just one, for maximising the utility of such questions as illustrated below:

Example

(i) List one characteristic to distinguish between plants and animals. (One value point)

(ii) List two common features and two contrasting features of plants and animals. (Four value points)

(d) Limit the scope of the answer to such an extent that it can be attempted within the time limit (1-5 mts.). It is better to restrict the response to a maximum of 4 credit points. For example, avoid a question like; 'List six characteristics of living organisms'. If however, you are interested to know more, frame a long-answer question or two short-answer questions.

(e) Use precise and unambiguous language to pinpoint clearly the specificity of the intended response to ensure better reliability of scoring. 'Write a short note on virus', is a bad example of short-answer question.

(f) Allocate marks according to the number of value points being tested. As far as possible marking for various value points may not be reduced below half a mark each to avoid difficulty in scoring. Moreover the marks allotted to the question should be divisible according to the number of intended value points. Allocation of two marks, for example, would create difficulty in scoring if there are three value points expected in the students' responses.

(g) Write model answers for each question as a complete answer bringing out clearly the value points which you would expect from the pupils. This helps to improve the language and definiteness of the scope of answer and also controls the reliability of marking.

5. Very Short-Answer Questions

In accordance with the mode of answering and questioning we may have the following types of very short-answer questions:

5.1. *Direct Question Form*

A straight question is put requiring a specific answer to be supplied by the students.

Example

(i) What is the technical term used for a whorl of petals in a typical flower?
(ii) Who invented the telephone?

5.2. *Selection Form*

A direct question is put with 2 or 3 choices designated in the question itself and the student is to write the answer after making a proper selection. The student may be asked to supply the answer or cross out what is not required.

Example

(i) Which of the two, 'go' or 'going' is a present participle?
(ii) The founder of Mughal empire was Akbar/Babar. (Delete which is not required)

5.3. *Completion Form*

A question is given in the form of an incomplete statement and the student is to fill in the appropriate word or phrase.

Example

(i) The first Russian to go into outer space was.....
(ii) Relationship between a carpel and the ovule is the same as between a stamen and the ....
(iii) In the life history of the frog the tadpole larva resembles the fish. The phenomenon indicates that.....

(Ontogeny repeats phyliogeny)

5.4. *Identification Variety*

In this case a student is asked to choose or underline a particular word with respect to the point of relevance indicated in the direction of the question.

Example

(i) In the following sentence underline the word which is an adjective.

I always prefer to buy a *black* pen.

(ii) Which of the two, Arunachal and Goa is a Union Territory?

5.5. *Recognition Variety*

In this case either a specimen, an apparatus, an object or sketches are given out of which one is to be recognised by careful observation.

Example

(i) Which of the given objects is a thistle funnel? (To be pointed)

(ii) Which of the given apparatuses is used for filteration of water? (To be selected)

(iii) Which of the sketches provided, is of Fahrenhiet thermometer? (To be ticked or underlined)

6. Suggestions for Construction of Very-Short-Answer Questions

Whichever form is used depending upon the teachers' preference and students' experience, care must be taken to restrict the response to one word, a phrase or at the most a sentence. It is essential because such questions are of the fixed-answer type and as such they should preferably be included in the same section of the question paper where objective type questions are put. Generally teachers find it convenient to put such

questions in a question form and students also respond favourably to such questions as they elicit direct responses. Much depends on the construction of the question itself to get the desired response. A few suggestions given below would be helpful to achieve this purpose:

(a) Frame the question in such a way that it has only *one correct answer*. This can be done by requiring a significant or key word as the expected response. Otherwise the item will not contribute to the objective intended because there could be many correct answers with reference to different synonyms or near synonyms.

Example

Poor : Who wrote the 'Origin of Species'? (Charles)

Improved: The name of the author of the 'Origin of Species' was.....

In the first case, the answers may be many e.g. an author, a biologist, a scientist, an Englishman, an English scientist etc. each one of which would be technically correct although the desired response is Charles Darwin.

(b) Use precise and unambiguous language to make the answer definite.

Example

Poor : An animal which eats the flesh of other animals is.....
(carnivorous)

Improved: An animal which eats the flesh of other animals is classified as.....

In the first question the answer may be, tiger, wolf, meat-eater etc. which is not the desired response while in the latter, the answer is pinpointed.

(c) Minimize the use of textbook phrases or stereotyped language in phrasing the questions. Such phrasing rewards and encourages rote memorisation which is not always associated with real understanding and often causes

ambiguity.

Example

Poor : For every action there is equal and.....
(opposite reaction)

Improved: A peace of stone when struck against a wall comes back due to Newton's.....law of motion.

(d) A direct question form is generally preferable to an incomplete statement especially in case of younger students.

Example

Poor : The capital of Sikkim is .....

Improved: Which is the capital of Sikkim?

(e) Avoid grammatical and other clues.

Example

Poor : The group of tissues form an.....(organ)

Improved: An organ is formed by a group of.....(tissues)

In the first case 'an' gives clue to the answer 'organ' which can be associated with the article placed before it.

(f) Avoid over-mutilated statements involving a number of blanks. When too many blanks are left in an incomplete statement, the meaning is almost lost.

Example

Poor : The.....by which.....is separated into various.....having different..... is known as ....

Improved: The process by which petroleum is separated into various components having different boiling points is known as.....
(fractional distillation)

(g) Avoid indefinite statements.

Example:
Poor : When was Mahatma Gandhi born?
Improved: In which year was Mahatma Gandhi born?

In the first case answer could be the date, month, year or even the 'same day' as Lal Bahadur Shastri.

(h) Blanks should be kept uniform in length to avoid any clue regarding the length of response expected.

(i) Blanks should be arranged in a manner convenient for scoring. It is better to have blanks towards the right hand side of the question or the statement.

(j) Before scoring prepare a key which contains all acceptable answers.

Example:
Which method is used for carrying ripe fruits to long distances?

In this case all possible responses like refrigeration, preservation by ice, cooling, etc. should be treated as correct.

This variety of short-answer question should be carefully framed to restrict the response to the desired extent, otherwise it leads to difficulty in scoring. As the response in such questions is just one word, a phrase or at the most a sentence, a very careful structuring of questions is needed in order to test effectively the objectives other than mere factual information.

This variety of short answer questions i.e. very short-answer type questions should be very carefully worded to restrict the response to the desired extent otherwise it may lead us to trouble while scoring such items. Moreover, unless special care is taken these questions tend to test merely recall of specifics, rather trivial details like names, dates, events, etc. As the responses in such questions is just one word, a phrase or a small sentence, a very careful structuring of such questions is needed in order to test effectively objectives other than mere factual information.

ANNEXURE

Objective-based Short-Answer Questions
For class VI-VIII

Concept

For germination of seeds, air, water, and suitable temperature are necessary.

Illustrative questions testing different abilities

1. Knowledge Objective

1.1. *Recalls*

List three conditions necessary for germination of seeds.

1.2. *Recognises*

In the absence of which of the following conditions can a seed germinate?

(i) Moisture, (ii) Temperature, (iii) Oxygen, (iv) Soil (V.S.A.)

2. Understanding Objective

2.1. *Translates*

Define the term 'germination' in your own words.

2.2. *Illustrates*

Give one example of a seed that requires more moisture for germination and one requiring comparatively less moisture.

2.3. *Identifies Relationship*

Why does a seed not germinate when sown too deep?

2.4. *Compares*

In what two respects does germination of a gram seed differ from that of a castor seed?

2.5. *Classifies*

Categorise the following seeds into two groups on the basis of their mode of germination.

A. Pea	C. Castor	E. Sun Flower
B. Gram	D. Maize	F. Onion

2.6. *Detects Errors*

Observe the given diagram—B of the experiment to show the conditions necessary for germination and point out the mistake if any. Suggest rectification also.

2.7. *Interprets*

What does this experiment demonstrate?

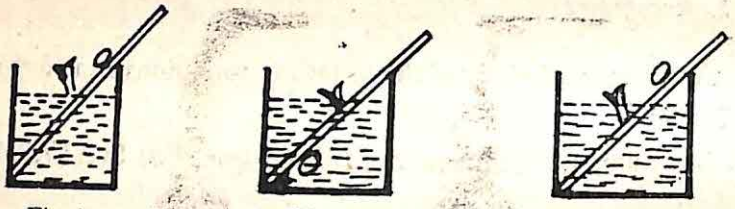


Fig-A Fig-B Fig-C

2.8. *Explains*

Why do we use three seeds instead of one in the above experiment?
(Figure—A Question 2.7)

3. Application Objective

3.1. *Analyses*

Some quality seeds of wheat were sown in two separate pots—A and B using the same quality seeds and same sample of moist soil and kept at the same place. Seeds in pot-A germinated but not in pot-B. Under what condition could this result be relevant?

3.2. *Hypothesises*

When mustard seeds are scattered on suitably prepared soil, they germinate but maize seeds do not. When sown a little deep under the same soil maize seeds germinate but not the mustard seeds. Propose a hypotheses to explain these observations.

3.3. *Suggests Procedure*

Bigger seeds required more moisture than smaller seeds for germination. Suggest an experimental procedure to test this hypothesis.

3.4. *Gives Reason*

Why do seeds generally not germinate when there is a heavy rainfall immediately after sowing?

3.5. *Draws Conclusion*

What conclusion can you draw from the observations of experiment shown in figure-C, (Question 2.7).

4. Skill Objective

4.1. *Draws Sketches*

Draw a labelled diagram to show the experimental set-up of the three bean experiment.

4.2. *Fits up Apparatus*

Examine the diagram given on the black board and fit up the apparatus to demonstrate the three bean experiment.

4.3. *Observes and Records*

Examine the three diagrams carefully. Note and record the difference in the conditions of germination available in the three cases

CHAPTER X

OBJECTIVE-TYPE ITEMS (MULTIPLE CHOICE TYPE)

1. Introduction

Of late objective-type items have become quite popular in India, partly because of their inherent objectivity in scoring but more so perhaps due to the increasing demand of such items in selection tests. Inspite of the fact that these objective-type tests are increasingly used in most of the entrance examinations for admission to medical, engineering, agriculture, commerce and other professional courses, the quality of such tests is far from satisfactory. Leaving aside a few specialists or trained evaluators, there is a dearth of good item writers. For classroom teachers, the use of such tests is still a wishful thinking because of two reasons. Firstly, the teacher by and large are not thoroughly trained in the technology of objective questions. Secondly, the use of such tests demand a lot of facilities like stencilling, duplicating of paper, stiching and administration which are generally not available. Nevertheless, once such tests are developed, cyclostyled or printed they may be used over a period of time as they can be kept confidential after every administration.

This chapter is devoted to one of the mostly commonly used farm of objective-type items, namely, the multiple choice type. Attempt is made that the study of this chapter enables the reader to

(a) describe the nature, scope and characteristics of various forms of multiple choice questions,
(b) construct various forms of multiple choice questions testing different abilities implied under various objectives,
(c) detect various tppes of clues and other defects in the given items of multiple choice variety and rectify them,
(d) identify the basis of improvement of defective items,
(e) explain the merit and limitations of different form of multiple choice items,
(f) apply the principles underlying in developing stem and responses while constructing multiple choice questions,
(g) judge the quality of a given item using prescribed criteria,
(h) appreciate the role of multiple choice items in improving the validity and reliability of the measuring instruments.

These tests overcome the criticism levelled at essay questions which have poor content-sampling, unreliable scoring, and are time consuming to score and amenable to bluffing and other extraneous factors. Objective-type items ensure more adequate coverage as many questions can be asked in the same testing time resulting in higher reliability and more content validity, besides ease of scoring and interpretation. Some critics, however, contend that objective type items encourage rote memory and emphasise testing of factual information because the higher mental processes cannot be effectively measured through such tests. This, of course, is an erroneous belief and a misconceived notion. Even if the scruitiny of such item-pools reveals poor kind of items, it only reflects our bankruptcy in framing higher level items, testing abilities like application, analysis, synthesis or evaluation rather than the inherent limitation of the objective-type items. Discussion in the pages that follow would amply support my contention that if the item writer understands the purpose and technology of an item he can, with practice, frame items which are really thought-questions requiring students' ability to think.

The term 'item' is usually preferred to 'question' in case of objective-type items. Since every question need not be in a question form i.e. followed by a sign of interrogation but can be set even in the form of a complete or incomplete statement,

it is more appropriate to use the term item in case of the objective-type question. What is important is that objective-type items must have a precisely-worded, predetermined correct response whatever instructional objective it tests or the format it takes. Thus scoring of such items can be done by any person without a knowledge of the subject. It may even be scored by a machine as it is done now in some of the public examiantions conducted by various boards of school education in India.

There are many forms of objective-type tests depending upon the type of stimulus given and the type of responses provided within the item. A detailed classification of objective-type items is given here-in-after.

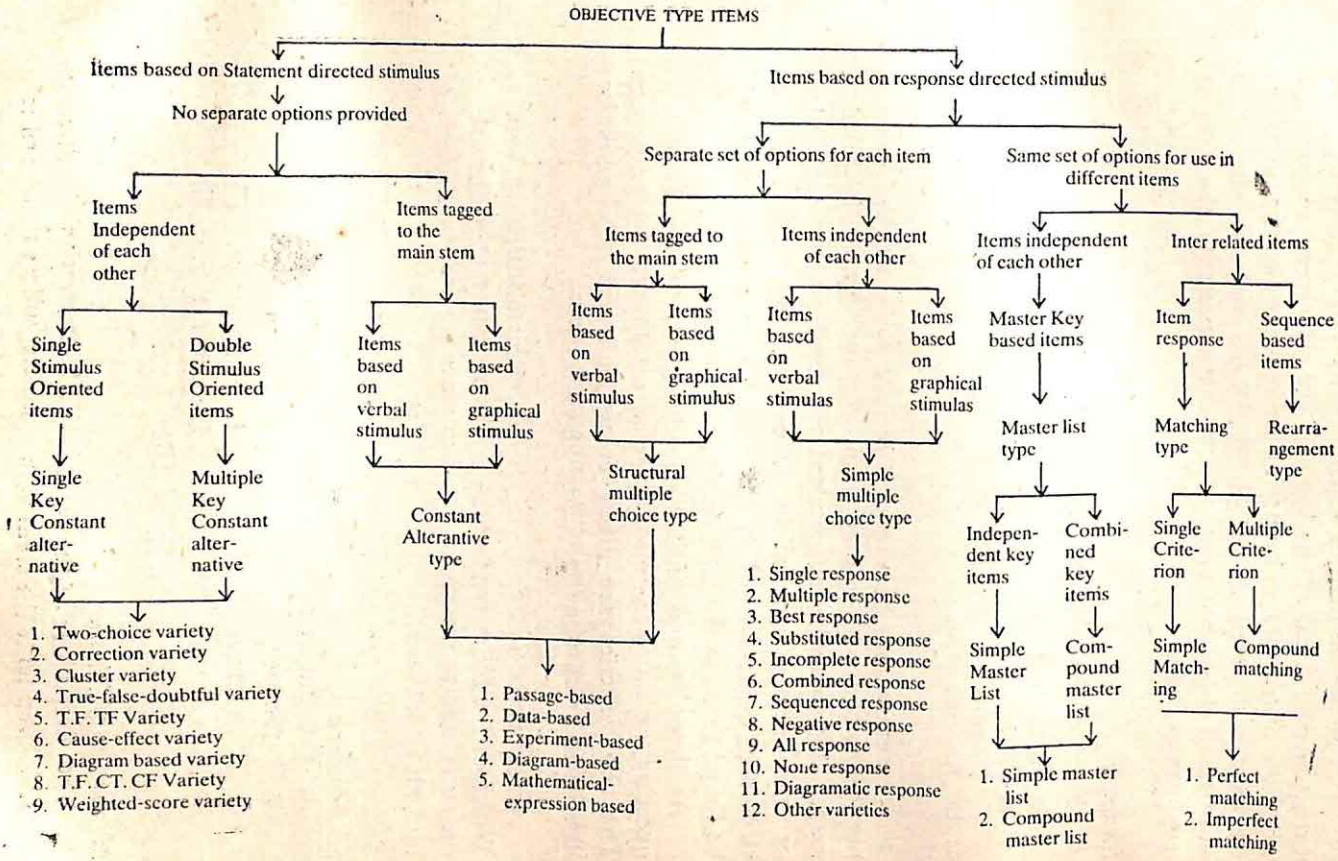
MULTIPLE-CHOICE ITEMS

As indicated in the classification of objective-type questions, the multiple-choice items are based on response-directed stimulus in which responses or options may be arranged in different ways. The choice of the correct answer among the given options or responses can be made from the independent set of responses given for each item or from the same set of responses given for different items. Accordingly, we have different varieties of multiple choice items in terms of design or format of the items as given in the classification flow-chart in the beginning.

Format

The format of multiple-choice items differs from type to type. Since basically all objective-type items (selection type) are multiple-choice items, the variation is in the manner in which the stimulus is provided and the responses arranged, as also the mode of responding or writing the answer. That is why the design of each type of multiple-choice item as also the characteristics and hints for construction are discussed separately. In this chapter we take the traditionally-used multiple choice variety.

An ordinary multiple-choice item consists of an incomplete statement or a question called the *stem* of the item followed usually by four or five suggested answers (choices, called the


OBJECTIVE TYPE ITEMS
Items based on Statement directed stimulus
No separate options provided
Items Independent of each other
Items tagged to the main stem
Single Stimulus Oriented items
Double Stimulus Oriented items
Single Key Constant alter-native
Multiple Key Constant alter-native
1. Two-choice variety
2. Correction variety
3. Cluster variety
4. True-false-doubtful variety
5. T.F. TF Variety
6. Cause-effect variety
7. Diagram based variety
8. T.F. CT. CF Variety
9. Weighted-score variety
Items based on verbal stimulus
Items based on graphical stimulus
Constant Alterantive type
1. Passage-based
2. Data-based
3. Experiment-based
4. Diagram-based
5. Mathematical-expression based
Items based on response directed stimulus
Separate set of options for each item
Same set of options for use in different items
Items tagged to the main stem
Items independent of each other
Items based on verbal stimulus
Items based on graphical stimulus
Structural multiple choice type
Items based on verbal stimulas
Items based on graphical stimulas
Simple multiple choice type
1. Single response
2. Multiple response
3. Best response
4. Substituted response
5. Incomplete response
6. Combined response
7. Sequenced response
8. Negative response
9. All response
10. None response
11. Diagramatic response
12. Other varieties
Items independent of each other
Inter related items
Master Key based items
Master list type
Independ-ent key items
Combi-ned key items
Simple Master List
Com-pound master list
1. Simple master list
2. Compound master list
Item response
Sequence based items
Matching type
Rearra-ngement type
Single Crite-rion
Multiple Crite-rion
Simple Match-ing
Compound matching
1. Perfect matching
2. Imperfect matching

responses or *options*) out of which usually one is correct. The correct answer is called the *key* and the incorrect answers are called the *distractors*, of course all being plausible. The responses are usually lettered as A, B, C, D, E, and the student is required to choose the correct letter that represents the correct response i.e. key.

Example

Which of the following towns is a state capital?] Stem

A. Allahabad
B. Patiala
C. Nagpur
D. Bhopal

options or responses — Distractors

] Key

A common multiple-choice item is usually written material but it may be based on a diagram, picture, graph, table or a map. Items may be independent or tagged to the same stem or introductory material in which case they are termed as structured questions. A more detailed description of these will be found towards the later part of this chapter.

As already stated, an item can be given in the form of a question or an incomplete statement followed by responses. The two varieties are given below followed by guidelines indicating the mechanics of writing such items.

Question form	*Incomplete statement form*
Which of the following instruments is used to measure altitude of a place?	For measuring the altitude of a place we use
Q. 1.	
A. Thermometer	A. thermometer. ☐
*B. Barometer	B. barometer. ⊕
C. Hydrometer	C. hydrometer. ☐
D. Lactometer	D. lactometer. ☐
1 (B)	

Mechanics of Writing

1. The first letter of the word of each alternative is written in small capital if the stem is incomplete, unless this word is a

proper noun.

2. Each alternative is followed by a full stop in case of the incomplete stem.
3. The first letter of the word of each alternative is written in capital when the stem is in a question form.
4. Write alternatives below the stem a little away (two spaces) from where the stem begins.
5. Underline or capitalise words like 'not', 'best', 'except', 'least' etc. when used in the item stem.
6. Use Arabic numerals for serialing the item.
7. Use capital A, B, C, D, and E for numbering the responses.
8. Specify the place either by a blank—— or a box □ or by a bracket ( ) for indicating the response, towards the right hand after the last alternative and number the bracket by the corresponding serial number of the item like 5 ( ). Sometimes it is advisable to put boxes or brackets against each of the options.
9. Mode of indicating the response e.g. (B) or ⊕ may be specified. For lower classes sometimes encircling of letter against correct answers is recommended.
10. For the same test keep the same number of options (4 or 5) for all the items. Four seems to be more practicable.

Arrangement of items

1. Group different types of items according to number of choices provided in different items.
2. Check whether correct responses in items are placed at random and not in some definite order to avoid clues.
3. See that all alternatives of the same question remain on the same page while writing for a test.
4. To save space, all alternatives may be written in a single row if they are small, or in two rows if needed, in classroom testing.
5. Provide common directions for similar types of items, e.g. single correct response, multiple correct response, re-arrangement type. . .etc.

POTENTIALITIES OF MULTIPLE CHOICE ITEMS

The very fact that in most selection tests only multiple-

choice tests are used, bears testimony to the inherent qualities of these items. Though it is not easy to frame good questions of this variety testing the higher abilities, yet with good practice and knowledge of the technology of developing such items, it should not be very difficult to frame questions testing almost all the cognitive abilities envisaged under Knowledge. Understanding and Application objectives as enunciated in the chapter on educational objectives. It is an erroneous belief that such items cannot be set to test higher level objectives. Following examples of objective-based items would amply support this view point. To avoid subject-based content difficulty, the content is taken from the pedagogy itself relating to the form of questions and the objectives.

ILLUSTRATIVE ITEMS TESTING DIFFERENT ABILITIES

1. Knowledge Objective

1.1. *Recognises/Recalls*

Which of the following forms of question is of the free response type?

*A. Completion type
B. Multiple choice
C. True-false type
D. Matching type

2. Understanding Objective

2.1. *Translates*

Which one of the following terms is closest in meaning to the term taxonomy?

A. Objective
B. Construction
*C. Classification
D. Evaluation

2.2. *Cites Examples*

Which of the following abilities illustrates the objective 'Comprehension' according to Bloom's taxonomy?

A. Comparison
B. Classification
C. Prediction
*D. Extrapolation

2.3. *Identifies Relationship*

If the number of questions in an objective-type test is reduced from 60 to 40 which of the following characteristics of the test would NOT be affected?

A. Validity
B. Reliability
C. Practicability
*D. Objectivity

2.4. *Compares*

Which of the following characteristics of a question is common to both essay and short-answer type questions?

A. Fixed response
B. Extended response
*C. Freedom of response
D. Selection of response

2.5. *Classifies*

Which of the following pairs of objectives is NOT grouped in order of hierarchy of Bloom's taxonomy?

A. Knowledge and Comprehension
B. Comprehension and Application
C. Analysis and Synthesis
*D. Analysis and Evaluation

2.6. *Detects Errors*

In terms of Bloom's taxonomy, in which of the following sets of abilities, arranged in order of complexity did a student make the mistake?

A. Knowledge of specific facts, principles, theories
*B. Interpretation, translation, extrapolation
C. Analysis of elements, relationships, organisational principles
D. Production of unique communication, plan, derivation of abstractions

2.7. *Interprets*

If a true-false test is to replace a multiple-choice test the number of items need to be greater in order to maintain almost the same

A. Level of difficulty
*B. Degree of reliability
C. Coefficient of validity
D. Extent of objectivity

2.8 *Explains*

In an essay-type test, we should not permit overall options (Any 5 out of 9 type) because that ensures much better

A. Content coverage
B. Coverage of objectives
C. Objectivity in scoring
*D. Comparison of students' performance

3. Application Objective

3.1. *Analyses*

In a multiple-choice test with four alternatives, the students can

attempt 25% items by guessing alone. Which one of the following assumptions the author is making?

A. That the students would make some informed guesses
*B. That the students would make blind guesses
C. That the students would make guesses in 25% questions
D. That 25% students would make blind guesses

3.2. *Hypothesises*

In Bloom's taxonomy, the educational objectives are arranged in increasing order of complexity (hierarchy) i.e. Comprehension is more complex than Knowledge; Application more complex than Comprehension and so on. Moreover, each higher-level objective subsumes the next lower level objective i.e. Evaluation includes Synthesis which includes Analysis that in turn includes Application and so on. From this we can hypothesise that factor analysis should exhibit

*A. 1 unique ability, 1 common ability and 4 group abilities
B. 2 unique abilities, 5 common abilities and 6 group abilities
C. 2 unique abilities, 4 common abilities and 5 group abilities
D. 1 unique ability, 6 common abilities and 1 group abilities

3.3. *Suggests New Procedures/Alternatives*

An objective test consisting of 50 True-False items was found to have a low reliability of 0.50. Which of the following measures is likely to increase its reliability the MOST?

A. Convert all the 50 items into completion type
*B. Convert all the 50 items into multiple-choice type
C. Increase the number of items to 60
D. Reduce the number of items to 40

3.4. *Gives Reasons/Establishes Relationships*

A blue-print of a question paper is a three dimensional chart in which distribution of questions and marks of various forms

of questions testing different objectives relating to different content area are shown. That is why a good blue print ensures better validity because of adequate sampling of

A. Content, elements, in each unit
B. Instructional objectives
C. Questions relevant to unit-objective
*D. All the above factors

3.5. *Concludes/Infers*

An 'Application' item measures 'Understanding' and 'Knowledge' objectives also, but an item testing Understanding does not test 'Application' but does test 'Knowledge' objective. A 'Knowledge' item neither tests the Understanding nor the Application objective. From this the best conclusion that can be drawn is that objectives are

A. Arranged in hierarchical order but are not cumulative
B. Cumulative in nature but not hierarchically arranged
*C. Arranged in hierarchy and are cumulative also
D. Neither hierarchical nor cumulative in nature

3.6. *Predicts*

Supposing a teacher teaches biology for three objectives namely, Knowledge, Understanding and Application. If in a unit test consisting of 50 multiple-choice items testing Knowledge and Understanding, he includes 5 more items based on the Application objective, what would happen to the validity and reliability of the unit test?

A. Validity would decrease but reliability would increase
B. Reliability would decrease but validity would increase
*C. Validity as well as reliability would increase
D. Reliability as well as validity would decrease

3.7. *Judges (evaluates)*

Essay type tests overrate the importance of 'How to say a

thing rather than having something to say'. This observation can be justified if we believe that essay type tests are

A. More reliable than objective-type tests because they ensure better coverage of content through long answer questions
*B. Less reliable than objective-type tests because validity of students' responses is affected by extraneous factors
C. More valid than objective-type tests because they ensure better the validity of responses
D. Less practicable than objective-type tests because they are difficult to score and interpret.

VARITIES OF MULTIPLE CHOICE ITEMS

1. Single-response Variety

This is the most commonly used form of multiple-choice items which are usable in classroom testing, public examinations and selection tests. These items have 4 or 5 alternatives with one and only one correct answer.

Example
Which of the following enzymes is produced from the salivery glands during digestion of food?

A. Trypsin
*B. Ptyalin
C. Pepsin
D. Erepsin

2. Multiple-response Variety

In this type, more than one answer may be correct. The choice of the correct answer is, therefore, not limited to one key but may be more than one. A specific direction is therefore necessary for the examinees requiring them to identify more than one Key. Such items are difficult to score and use in public examinations but they are very good for use in unit testing to know students' real understanding of the content-

elements and provide for good training in improving their discriminating power.

Example
Which of the following plants can be classified as hydrophytes?

A. Cuscuta
*B. Lotus
C. Calotropis
D. Hydrilla

3. Best-response Variety

In this case more than one could be the correct answer or all of them may be correct. But the one that describes or explains best the concept, principle or the idea, is considered the correct answer. Such questions test students' ability to discriminate between closely-related concepts. In such questions, the direction must be clear and the word 'best' be underlined or capitalised as shown below:

Example
Which of the following chemical equations explains BEST the mechanism of photosynthesis?

A. $6CO_2+6H_2O=C_6H_{12}O_6+6O_2$
B. $6CO_2+12H_2O=C_6H_{12}O_6+6H_2O+6O_2$
C. $6CO_2+12H_2O^{16}=C_6H_{12}O_6+6H_2O+6O_2{}'^{6}$
*D. $6CO_2+12H_2O^{18}=C_6H_{12}O_6+6H_2O+6O_2{}^{18}$

4. Substituted-response Variety

In this type the blanks are left to be filled-in by the examinee on the basis of options or responses provided after the statement. Such items are useful when the content of learning demands the use of the same type of responses to every teaching point. In such cases, the same alternatives may be used for filling up the blanks by substituting the relevant alternative.

Example
Deficiency of vitamin—A causes (*Night-blindness*)
While lack of vitamin—C causes (*Scurvy*)

A. Anemia
B. Beriberi
C. Scurvy
*D. Night-blindness

5. Incomplete-response Variety

In this type, the stem of the item is followed by the four or five options which are given in the form of incomplete responses. The testee is given hints in the form of limits of letters, incomplete word or a single letter. Such items can only be used profitably in testing vocabulary, terminology, spellings etc.

Example
(a) One who is deathless, everlasting and indestructible is called:

A. o *B. i C. d D. s (*immortal)

(b) An oval figure is termed as

A. Cu..... B. Cy..... C. ro..... *D. el..... (elliptical)

(c) The spelling of the word 'rec-pt' will be correct if the two blanks are filled up using.

*A. ei B. ee C. ie D. ii

6. Combined-response Variety

Sometimes when one is interested in judging more thoroughly about a concept, a combination of responses may be used as key in place of the single-response key.

Example (uncombined key)
Lack of vitamin A causes

Example (combined key)
Source of Vitamin A is

A. anaemia
B. beribery
C. scurvy
*D. night blindness

a. Liver
b. Butter
c. Milk
d. Grains

A. a only
B. a+b
*C. a+b+c
D. a+b+c+d

7. Sequenced-response Variety

This type involves responses which depict some sequence of items, objects, procedure, events etc. Such items ensure lot of coverage of content besides testing students' ability to identify relationships and classify facts, concepts, events, etc. Since such items are based on a very comprehensive and broad chunk of syllabus, they are very economical per unit testing time. However, such items can be framed only on such content areas as involve sequential matter.

Example

Which of the following sequences respresents the evolutionary development of animals?

A. Snake—→Frog—→Pigeon—→Cow
B. Frog—→Pigeon—→Cow—→Snake
C. Pigeon—→Cow—→Snake—→Forg
*D. Frog—→Snake—→Pigeon—→Cow

8. Negative-response Variety

Sometimes it is difficult to find four homogeneous alternatives but three are possible. In order to use these three, the item stem is made negative in order to accommodate the fourth alternative, which is not related, as the desired key.

Example

Which of the following is NOT necessary for germination of a seed?

A. Air
B. Water
*C. Soil
D. Warmth

9. All or None-response Variety

Sometimes it is difficult to write an item because all the responses are good as an answer or key but one finds difficulty in identifying distractors. In such cases one may use the fourth (last) option as 'all the above'. Likewise, if one is not able to find even a single good alternative relevant to the situation depicted in the stem, then relevant distractors are chosen as alternatives and the fourth (last) option may be listed as 'None of the above' as illustrated below:

Example
which among the following phenomena is based on the principle of air pressure?

A. Working of water pump
B. Filling in of ink in a fountain pen
C. Working of a siphon
*D. All the above

Which of the following phenomena reflects the principle of Archimedes?

A. Rising of water in a straw pipe when sucked in
B. Working of a siphon
C. Working of a water pump
*D. None of the above

10. Multiple Selection/Completion Variety

In this type, four choices are given as A, B, C & D and the key may comprise more than one of these responses. The student is required to find the correct answer in terms of the stem of the item, the answer being only A or B or C or D, each

having more than one response combined under each. These items can be used when we want to test students' ability to cite examples, compare, discriminate or analyse a situation involving allied concepts or ideas.

Example
Which of the following biologists has worked on the theory of organic evolution?

1. Darwin 2. Lemarck 3. Mendel 4. Wallace.

Select:
A. if 1, 2 and 3 are correct
B. if 1, 3 and 4 are correct
C. if 2, 3 and 4 are correct
*D. if 1, 2 and 4 are correct

11. Worst-answer Variety

In such items, all the four or five alternatives are relevant except the one which is the worst. Students are asked to choose the worst and write its number in the space on the right. Such items can be used to test students' ability to detect errors, irrelevance and incongruency in statements.

Example
A rhombus is a geometrical figure in which

A. all sides are equal
*B. all angles are equal
C. opposite sides are equal
D. opposite angles are equal

12. Most-inclusive Variety

Four or five words are given in a set, one of which includes the other three or four. A Student is to select the word which is the most inclusive i. e. covering the others and write its serial number in the space provided on the right. Such items can be

used to test vocabulary, terminology, relationship, similarities etc.

Example

(a) A. Beribery B. Typhoid C. Disease D. Malaria *C*
(b) A. Spirogyra B. Ulothrix C. Chlamydomonas D. Algae *D*

13. Most-dissimilar Variety

A set of four words or terms are given, one of which does not belong to the other three. Students are to choose the most dissimilar one and write its serial on the space provided on the right. This type is useful to test students' ability to identify relationships and classify on the basis of implicit criterion.

Example

1. A. Cuttle fish B. Star fish C. Shark fish D. Devil Fish *C*
2. A. Jahangir B. Altmash C. Akbar D. Babar *B*
3. A. Hard B. Last C. Fast D. Cast *D*

14. Cause-effect Variety

In this case events representing causes (or effects) are given, out of which one is the effect (or cause) respectively. Students are required to select the effect (or cause) and write its serial in the space provided on the right. Such items can be used to emphasise (in instruction) students' ability to establish the cause and effect relationship.

Example

1. Which of the following events is an effect of the other three causes of the downfall of Mughal empire?

   A. Weak successors after Aurangzeb
   B. Religious policy of Aurangzeb
   C. Faction fighting among nobles
   *D. Rise of independent kingdoms

2. Which of the following events is the cause of the other three effects?

A. Rise of independent states
*B. Lack of political insight of Aurangzeb
C. Consolidation of Marathas in Deccan
D. Invasion of Nadir Shah to loot Delhi

15. Analogy-type Items

In this case three words, terms or concepts are given, showing a relationship between the first and the second on the basis of which the fourth word is to be identified out of 4 or 5 given as alternatives.

Example

Identify the fourth term in the following items and indicate by encircling the serial number of the correct response.

(i) Sun : : Mercury : : Earth:————————

A. Jupiter
B. Mars
*C. Moon
D. Venus

(ii) Gujarat: Ahmedabad : : Manipur:———— ——

A. Dispur
B. Shillong
*C. Imphal
D. Kohima

16. Multiple Completion/Selection Variety

In this case, the stem of the item is followed by options having one or more than one correct response. A student is required to select the correct response or responses. However, to facilitate scoring, the responses are coded for various combi-

nation of responses so that each item has only one lettered option as its key.

Example
Indian Prime Ministers since 1947 include

I. Lal Bahadur Shastri
II. Radhakrishnan
III. Charan Singh
IV. Morarji Desai

	Select
A. I and II only	A. if 1 & 2 are correct
B. II, III and IV only OR	B. if 2, 3 & 4 are correct
*C. I, III and IV only	*C. if 1, 3 & 4 are correct
D. I, II, III and IV	D. if 1, 2, 3 & 4 are correct

OR when each item consists of 4 options arising from the *three* numbered responses the code may be:

Select A. if 1, 2 and 3 are correct
B. if 1 and 2 only are correct
C. if 3 only is correct
D. if some other single response or combination of responses is correct

Here one code is left unspecified (D) and the student must show whether each option considered on its own is or is not the correct answer. It is therefore a more thorough probing, allowing for the least guess work. Such items lead to muddy measurement sometimes if the aim is not clear. If the intention is to find out whether a student knows which of the possible answers *are* correct and which *are not* correct then such items can be used more profitably.

17. Assertion/Reasoning Variety

Such items are based on two statements, the first being coded as 'assertion' and the second as 'reason' and linked by word '*because*'. In fact such items provide a technique for converting true-false items into the multiple-choice format.

Example

An iron needle sinks in water while a toy ship having the same weight floats. (Statement 1)

because

the volume of water displaced by the needle is more than the weight of the needle. (Statement 2)

Select A. if both the statements are correct and the second is the explanation of the first.

*B. if both statements are correct but the second is NOT the explanation of the first.

C. if the first statement is correct but the second is incorrect.

D. if the first statement is incorrect but the second is correct.

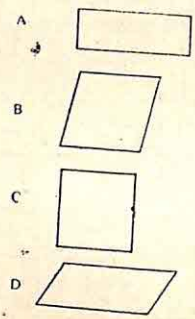
Such questions more often test students' understanding of concepts involving higher level abilities and not mere recall.

18. Diagrammatic-response Variety

In this case, instead of verbal responses, diagrams may be given as responses (options) and students are required to select the correct diagram (key) out of 4 and 5 given. Such items are useful to test perceptual ability along with functional understanding.

Example

In which of the following diagrams are all the sides as well as the angles equal?



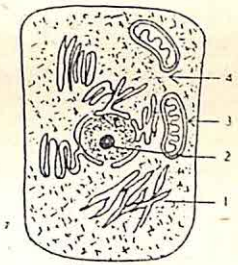
19. Diagram-based Response Variety

In this case a diagram, table, chart or a picture is given and verbal responses are provided based on the given diagram. Such items can be used to test various abilities depending upon the nature of graphic stimulus used.

Example

In the given electronic diagram of an animal cell the organelle which is concerned with respiration is indicated by the label

A. 1 B. 2 *C. 3 D. 4



Animal Cell

SUGGESTIONS FOR CONSTRUCTING MULTIPLE-CHOICE ITEMS

A. General

(i) *Develop multiple choice items on the basis of independent meaningful and important ideas.*

Example

Poor: Which of the following formulae is used to estimate the reliability of a test by the split-half method?

A. Spearman-Brown
B. Chronbach
C. Kudar-Richardson
D. Gutman

Better: Spearman-Brown formulae is useful in estimating the reliability when

A. the same test has been given twice.

B. the only available data are the number of items in the test and the mean and standard deviation of the scores.
C. one knows the number of items in the test.
*D. a coefficient of correlation between scores on odd and even numbered items in a test has been calculated

(ii) *Item should be strictly true*

Example

Poor: Why is blood plasma often preferred to whole blood for transfusion?

A. Whole blood may carry disease germs.
*B. Whole blood must be "typed" to match the blood of the patient.
C. Plasma can be prepared synthetically.
D. Plasma contains more disease finding white corpuscles than whole blood.

It is not correct the *blood plasma* is preferred to whole blood, because if whole blood of the proper type is available, it is usually preferred. The stem should ask '*what advantage does blood plasma have over whole blood*' because it is true that plasma need not be typed.

(iii) *Frame questions using definite and concise language to get the intended response without further qualifications*

Example

Poor: Which one of the following is the best source of heat for home use?

*A. Coal B. Electricity C. Gas D. Oil

Better: In the northern districts of Himachal Pradesh which one of the following is the most economical source of heat for home use?

*A. Coal B. Electricity C. Gas D. Oil

(iv) *Language of the item should be comprehensible*

Example

Poor: A chaotic situation means

A. Asymptotic *B Confused C. Gauche
D. Permutable

Better: A. Pithy *B. Confused C. Tactless
D. Interchangeable

(v) *Avoid basing items on unique organisation*

Example

Poor: The first characteristic of a living organism is its ability to

A. respire B. reproduce C. grow D. move

Better: The first Mughal King was

A. Akbar *B. Babar C. Jahangir D. Humayun

(vi) *As far as possible write an item in a positive question form rather than in a negative and incomplete form*

(vii) *Avoid interlocking of items*

Example

Poor: A seed can germinate in the absence of

A. Air *B. Light C. Moisture D. Warmth

Better: Air is essential for germination of a seed because it provides to the growing embryo

*A. Oxygen B. Carbon dioxide C. Nitrogen
D. Hydrogen

If both these items are used in the same test, the second one provides a clue to the first.

B. Stem of the Item

The following criteria may be observed for writing the stem part of the item:

(i) *Task must be set in the stem itself*
Poor: Essay type questions

*A. have full freedom of response.
B. are free from subjectivity in scoring.
C. are more reliable than objective type.
D. are easy to grade.

Better: Essay type questions *are characterised* by

A. ... B. ... C. ... D ...

(ii) *Pose a single central problem as far as possible*
Poor: *Scurvy and night-blindness* are caused due to the deficiency of

*1. Vitamin-A 2. Vitamin-B
*3. Vitamin-C 4. Vitamin-D

Better: Scurvy and night-blindness are caused due to the deficiency of vitamins

1. A and B 2. B and C 3. C and A 4. C and D

(iii) *Include as much part of the item in the stem as possible to reduce the reading load* (5)
Poor: Which is the best definition for a vein?

A. A blood vessel that carries blue-blood.
*B. A blood vessel that carries blood going to the heart.
C. A blood vessel that carries impure blood.
D. A blood vessel that carries blood away from the heart.

Better: A vein is a blood vessel that carries

A. blue-blood
B. impure blood
*C. blood going to the heart
D. blood away from the heart

(iv) *Avoid using a negative stem as far as possible*

Poor: In the definition of a mineral which of the following properties is *not* relevant?

A. It is produced by geologic processes.
B. It has distinctive physical properties.
C. It contains one or more elements.
*D. Its chemical composition is variable.

Better: In the definition of a mineral which of the following properties is relevant?

A. It is produced by the sedimentary process.
B. It has distinctive chemical properties.
C. It contains two or more elements.
*D. It has a varied chemical composition.

(v) *Avoid double negatives in the stem*

Poor: Questions which are *not* of the supply type *cannot* be classified as

A. True-false type
B. Multiple choice type
*C. Essay type
D. Matching type

(vi) *Avoid window-dressing* (*irrelevant material in the stem*)

Poor: Gill was suffering from night-blindness. His family doctor advised him to take more milk daily. This was due to fact that milk contains

*1. Vitamin-A 2. Vitamin-B 3. Vitamin-C 4. Vitamin-D

Better: Milk is useful to a patient of night-blindness because it

provides.

*1. Vitamin-A 2. Vitamin-B 3. Vitamin-C 4. Vitamin-D

(vii) *Avoid stereotypes in the stem*

Poor: For every action there is an equal and opposite reaction. Which of Newton's Laws does this statement refer to?

A. First B. Second C. Third D. None of these.

Better: When a ball is kicked against a wall it comes back immediately. Which of the Newton's Laws does this example illustrate?

A. First B. Second C. Third D. None of these

(viii) *Avoid asking for opinion instead of intended outcome*

Poor: What do you consider is the most appropriate form of questions for testing student's ability to comprehend a passage?

*A. Short-answer type
B. Multiple-choice type
C. Matching type
D. True-false type

Better: Which of the following forms of questions is most suitable for testing students' ability to comprehend a passage?

*A. Short-answer type
B. Multiple-choice type
C. Matching-type
D. True-false type

(ix) *Provide for all necessary qualifications in the stem*

Poor: What change would occur in the composition of air in a room in which green plants are growing?

A. Carbon dioxide increases and Oxygen decreases.
*B. Oxygen increases and Carbon dioxide decreases.
C. Both Carbon dioxide and Oxygen decrease.
D. Both Carbon dioxide and Oxygen increase.

Better: What change would occur in the composition of air in a *lighted, air tight* room in which the only living things growing are green plants.

A. ... B. ... C. ... D. ... (6)

(x) *Avoid instructional aids*

Poor: Cells combine to form tissues and tissues make up an organ. The relationship between a tissue and an organ is similar to the one between

A. Fats and carbohydrates
*B. Organ and the system
C. Liver hand the gall bladder
D. Ectoderm and endoderm

Better: The relationship between a tissue and an organ is similar to the one between

A. ... B. ... C. ... D. ...

C. Responses

(i) *Provide for 3 to 5 alternatives*

Three to five alternatives depending upon the nature of the teaching/learning point and maturity level of the pupils may be used. Less choice for younger ones and more for the test sophisticated may be useful. Well chosen three choices are quite usable in classroom testing, if four choices are not easily available.

(ii) *Item should admit only one correct or best answer*

Poor: Fish can be preserved in salt water because

A. salt acts as a poison for bacteria.
*B. bacteria cannot withstand the osmotic action of salt.
C. salt alters the chemical constituents of the food.
D. salt protects the fish from contact with air.

In this item responses A and B both could be judged as correct as response-B, Simply explains the response-A which also is correct.

Better: Fish can be preserved in salt water because

A. bacteria can survive more easily in salt water.
B. bacteria cannot withstand osmotic action of salt.
C. salt alters the chemical constituents of the food.
D. salt protects the fish from contact with air. (7)

(iii) *Avoid clues of various types*

(a) *Format clue* (wording of responses)
In the example (Poor) given above under (ii) the difference in use of the first word provides clue to the designated response i.e. 'B' because in all the three distractors the response begins with salt while in response 'B' which happens to be the correct answer the first word is 'bacteria', quite different from others.

(b) *Associational clue*
Poor: Some of the observable differences between organisms of the same kind is called

A. heredity B. adaptation
C. variation D. heterophylly

In this case 'C' has a verbal association with the word 'difference' as both variation and difference are synonymous and a student can guess the correct answer simply by this clue.

(c) *Grammatical clue*
Poor: Which of the following plants bears naked seeds?

A. Grasses B. Lilies *C. Pines D. Ferns

In this case the 'key' is in the singular and distractors are presented in the plural form.

Better: A. Grasses B. Lilies C. Pines D. Ferns

Poor: He ran *fast* on seeing the dog. In this sentence the word 'fast' is used as

A. a verb B. a noun
C. a preposition *D. an adverb

In this case the article 'an' in the key provides a clue for the correct answer.

Better: A. a verb B. a noun C. an adjective D. an adverb

(d) *Verbal clue by repetition of a word from the stem*

Poor: The relationship between a cell and a tissue is almost similar to the one between

A. Cell and the system
*B. Tissue and the organ
C. Liver and the gall bladder
D. Ectoderm and the endoderm

In this item the word 'tissue' provides a clue due to verbal association with the word 'tissue' in the stem.

Better: Replace the response—B by 'Organ' and the 'System'.

(e) *Response length clue*

Poor: A seed can germinate in the absence of

A. Moisture
B. Oxygen
C. Temperature
*D. Amount of light

In this case the length of response 'D', which happens to be the key, gives a clue.

Better: D. light.

(f) *Sentence completion clue*

Poor: When dilute sulphuric acts on Zinc

A. Carbondioxide
B. Sulphurdioxide
C. Nitrogen
D. Hydrogen is formed

Here a clue is given in 'D' by completing the sentence.

Better: Which gas is evolved when dilute sulphuric acid acts on Zinc?

A. Carbondioxide
B. Sulphurdioxide
C. Nitrogen
D. Hydrogen

(g) *Clue by determinders like 'never', 'always', 'seldom', generally*

Poor: Lack of vitamin-C in diet results in

A. Nervous disorder
B. Night-blindness
C. Beribery
*D. Bad gums generally

The word 'generally' provides the clue to the answer.

Better: Lack of vitamin-C in diet results generally in

A. Nervous disorder
B. Night-blindness
C. Beribery
*D. Bad gums

(iv) *Use heterogeneous responses or homogeneous responses depending upon the intended difficulty level*

Heterogeneous responses	*Homogeneous responses*
Cuscuta is an example of	Cuscuta is an example of
A. an angiosperm	A. a hydrophyte
B. a gymnosperm	B. a saprophyte
C. a fern	C. an epiphyte
*D. a parasite	*D. a parasite

(v) *Use only plausible distractors*

Poor: Which of the following types of questions ensures the LEAST scoring uniformity?

Poor	*Better*
A. Multiple choice type	A. Essay type restricted response
B. True-false type	B. Short answer type
C. Matching type	C. Very short answer type
*D. Essay type	*D. Essay type-extended response

(vi) *Responses should be as brief as possible*

Poor: According to Bloom's taxonomy, the comprehension objective means

A. the ability to recall, translate and interpret
B. the ability to translate, analyse and interpret
C. the ability to interpret, extrapolate and evaluate
*D. the ability to translate, interpret and extrapolate

Better: According to Bloom's taxonomy, the ability to translate, interpret and extrapolate is covered under the objective

A. Knowledge
*B. Comprehension
C. Application
D. Analysis

(vii) *Avoid use of stereotypes except to misguide the rote learner*

Poor: In the life history of the frog the tadpole resembles a fish. This phenomenon illustrates that

*A. ontogeny repeats phyllogeny (poor use)
B. fishes have evolved from amphibians
C. toads and frogs have arisen simultaneously
D. fishes and frogs have evolved from a common ancestor

Good: The greatest improvement in live stock has been brought by

*A. selective breeding
B. organic evolution
C. environmental influence
D. survival of the fittest (good use)

(viii) *Avoid lifting a statement from textbooks*

Poor: How adequately the content of the test samples the domain of the subject matter about which inference is to be made, is termed as

*A. content validity
B. concurrent validity
C. predictive validity
D. construct validity

Better: A geography unit-test which is based on the content elements of that unit and instructional objectives indicated in its blue print reflects its

*A. content validity
B. concurrent validity
C. predictive validity
D. construct validity

(ix) *Write the responses in sequential order if one exists*

Poor: The population of India is about

A. 400 millions B. 550 millions
*C. 650 millions D. 750 millions

Better: If the radius of earth is increased by 5 feet its circumference at the equator would be increased by about

A. 6 feet *B. 31 feet C. 75 feet D. 300 feet (8)

(x) *Use 'all of these' or 'none of these' appropriately and sparingly*

When used make them correct answers as well as distractors as frequently as other distractors.

Poor: (i) Which of the following terms is closest in meaning to 'growth'?

A. Maturation
B. Learning
C. Development
D. None of these (Poor)

(ii) Mughal empire can be regarded as

A. controlled democracy
B. benevolent monarchy
C. parliamentary democracy
D. all of these (Poor)

Good: Which of the following words is mis-spelt?

A. Receipt
B. Belief
C. Perceive
D. None of these (good)

Mughal empire can be regarded as

A. controlled democracy
B. parliamentary democracy
C. presidential form of government
D. None of the above types (good)

(xi) *Ensure that all responses are independent of each other*

Poor: Muslim population in India is *more* than

A. 70 millions B. 100 millions
C. 150 millions D. 200 millions

In this case options overlap and the correct answer could be A and B both as in both cases the population will be more.

Better: Muslim population in India is *about*

A. 70 millions B. 100 millions
C. 150 millions D. 200 millions

(xii) *Avoid non-parallel responses*

Sometimes one of the four or five responses is not parallel to the other and becomes functionless as illustrated below:

Which of the following organisms does not have the capacity to manufacture its own food?

Poor	*Better*
A. A fern	A. A fern
B. A moss	B. A moss
C. An alga	C. An alga
*D. An animal (not parallel)	D. A fungus (parallel)

(xiii) *Use qualitative scale of responses when the answer varies from complete establishment to complete indefiniteness*

Desirable

Quite a few cases of dengu have been attributed to virus. What was the status of this idea in 1960?

A. This was well-established by medical sciences
B. It was a controversial matter and the evidence was inconclusive
C. This was completely rejected on the basis of sufficient evidence
*D. This was a recent development and needs further evidence and investigation

(xiv) *Combine two elements to give four responses when more coverage is desired or all the responses are not possible*

Desirable

If an essay test of six questions is replaced by three essay-type questions and ten short-answer type questions, it is likely to improve

*A. Validity as well as reliability
B. Validity but not the reliability
C. Reliability but not he validity
D. Neither validity nor reliability

(xv) *Avoid placing items in tandem* (Not desirable)
Sometimes the responses are placed in tandem i.e. one after the other to save space but this makes the student much more difficult to compare different responses in the process of its selection.

Example
The balance-sheet report of Gill's company would reveal

A. company's profit for the previous fiscal year
B. the amount of income tax paid
C. the amount of sales for the period and
*D. the amount of money owed to its creditors

(xvi) *Provide alternative definitions rather than words as alternatives while testing understanding of definitions*

Not desirable
From the lakes, rivers and oceans, water gets into the air by a process called

A. Filteration B. Condensation *C. Evaporation D. Distillation

Desirable
Evaporation is a process by which

A. vapours turn into liquids
B. liquids turn into solids
*C. liquids turn into vapours
D. solids dissolve into liquids

CHAPTER XI

OBJECTIVE-TYPE ITEM (OTHER VARIETIES)

1. Constant Alternatives Type
(True-False and allied varieties)

1.1. *Format*

The students' are required to select the answer from two or more choices or alternatives which remain the same for a whole series of items. These items are usually in the form of statements concerning whose truth or falsity the pupil is to make a judgement. In certain varieties, instead of a statement, a direct question may be asked for pupils to respond in a Yes-No form as given below:

Is lotus a water plant?	Yes-No
Lotus is an aquatic plant.	True-False

Responses of this type may take the form of True-False, Yes-No, Agree-Disagree, Synonym-Antonym etc. The statement is preceded by a clear direction regarding the stipulation of responses and mode of responding.

In continuous with the previous chapter on objective-type questions this chapter deals with other varieties of questions like those of constant alternatives, matching type, rearrangement-type, pictorial and structured questions. It is intended that the reader after going through this chapter will be able to:

(a) recognise the format, design and varieties of three questions,
(b) identify examples and non-examples of different varieties of such questions,
(c) apply the basic principles of construction of these question in framing questions of these types,
(d) describe the nature, scope, and limitations of these questions,
(e) select and use the relevant variety of items for testing the intended abilities, and
(f) suggest improvement in a given set of defective items.

The following five major varieties other than multiple choice type are described here-in-after.

1.2. *Forms of Constant Alternatives*

(a) *True-False variety*

(i) Temperature remaining the same the pressure in a gas varies directly proportional to its volume. *T*—F
(ii) Star fish belongs to the class, Pisces. T—*F*

(b) *Right-Wrong variety*

(i) Pinto ran fast in the 100 meter race. *R*—W
(ii) Ahmad as a fast runner of 100 meter race. R—*W*

(c) *Agree-Disagree variety*

(i) Religious dogmas lead to misinterpretation of religion. *A*—D
(ii) Every religion teaches hatred for other religions. A—*D*

(d) *Yes-No variety*

(i) Does a water pump work on the principle of

water pressure? *Yes*—No
(ii) Is Cinchona a medicinal plant? *Yes*—No

(e) *Synonym-Antonym variety*
Underline 'S' if the sentence reflects the use of two words similar in meaning and 'A' if used for opposite meaning.

(i) Delhi weather during May is very uncomfortable and uneasy. *S*—A
(ii) Hard water is not fit for drinking but soft water is. S—*A*

(f) *Correction variety*
Underline the false part of the statement and write the correction in the bracket provided against each statement.

(i) Water consists of two parts of hydrogen and *two parts* of oxygen by volume. (one part)
(ii) Water is a chemical *element.* (Compound)

(g) *True-False—Doubtful variety* (T.F.O.)
Underline T if true, F if false and O if the statement reflects an opinion.

(i) Mammals suckle their young. *T*—F—O
(ii) Earth revolves round the moon. T—*F*—O
(iii) There are no plants on Mars. T—F—*O*

(h) *Cluster variety*
In this case a complete stem followed by a number of statements are given for students to mark true or false.

(i) The volume of a mass of gas tends to
(a) increase with increase in temperature. *T*—F
(b) increase with increase in pressure. T—*F*
(c) decrease with increase in pressure. *T*—F
(d) decrease with increase in temperature. T—*F*

(i) *T-F—TF variety*
Against each statement underline T if the statement is

true under all circumstances and F if it is false under all circumstances and TF if it is true under certain circumstances and false under other circumstances.

(i) Sum of the angles of a triangle is equal to two right angles. *T*—F—TF
(ii) Respiration takes place in green cells only. T—*F*—TF
(iii) Living cells can manufacture food. T—F—TF

(j) *Converse True-False variety* (T.F. C.T. C.F.)
Underline T if the statement is true, F if it is false, also mark C.T. if converse is true and C.F. if converse is false.

(i) A square is a quadrilateral. *T*—*F*—CT—*CF*
(ii) Every parallelogram is a rhombus. *T*—F—*CT*—CF
(iii) An equilateral triangle is also equiangular. *T*—F—*CT*—CF
(iv) All parasites are animals. *T*—F—CT—CF

(k) *Cause-effect relationship variety*
Certain words are italicised in the following statements. Underline the word 'cause' if they reflect cause and underline 'effect' if they are effects of certain causes.

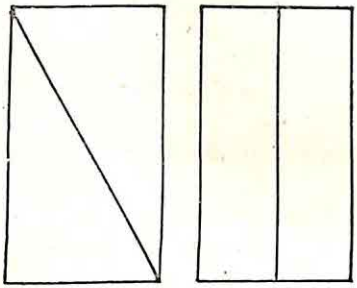
(i) *Tides* are formed during full moon. Cause—*Effect*
(ii) Due to absence of *Chlorophyll* a fungus cannot manufacture its food. *Cause*—Effect
(iii) Leaves of certain plants become yellowish in the absence of *light*. Cause—Effect
(iv) Aurangzeb's religious policy was responsible for the *downfall* of Mughal Empire. Cause—*Effect*

(l) *Qualified True-False variety*
Students are required to judge whether the statement is true or false and whether the item can be made true by qualifying the statements by adding some more information. In the following statements adding qualifications to make the item completely true if needed. Underline R or W.

(i) All leaves can manufacture food.	Add: 'green' before 'leaves' (*R*—W)
(ii) Under the Indian constitution the President can dismiss a Cabinet Minister.	Add: on the advice of the Prime Minister (R—W)
(iii) A speaker of a legislative assembly does not belong to any political party.	No qualification required (R—W)

(m) *Diagram-based variety*

In certain cases diagrams can be given and students may be required to make small judgements after observing the diagrams.



Observe the two diagrams of Door—A and Door—B with two iron bars joined in the frame.

(i) Door—A will become stronger than Door B T—*F*

(ii) Door—A will require more bar length than Door—B *T*—F

1.3. *Guidelines for Construction*

Following examples would illustrate different suggestions which may be considered useful for constructing true-false items:

(a) Item should be based on some *significant* and useful propositions:

Poor : Relationship between a parasite and the host plant is important.

Better: A parasite gets its food from the host plant.

(b) Statement should be made *absolutely true* or *absolutely false* by seeking experts' unanimity.

Poor : All acids are liquids.
Better: All acids turn blue litmus red.

(c) Necessary background or *qualification* must be incorporated in the statement to pinpoint the intended response.

Poor : Objective-type questions are better than essay-type questions.
Better: From the point of view of objectivity in scoring, the objective-type questions are better than essay-type questions.

(d) Avoid *half-true* and *half-false* statements.

Poor : The sun is a star while the Earth is a satellite.
Better: The sun is a star while the Earth is a planet.

(e) Avoid specific determiners like 'always', 'frequent', 'never' etc. as they provide clues.

Poor : Scoring objectivity in essay type questions has always been questionned.
Better: Grading of essay-type questions is questioned for subjectivity in scoring.

(f) Express *single* idea in each statement.

Poor : Solar and Lunar eclipses are formed when the Moon comes between the Sun and the Earth or the Earth comes between the Sun and the Moon.
Better: A Lunar eclipse is formed when the Earth comes between the Sun and the Moon.

(g) Use unambiguous and precise language to make the meaning clear.

Poor : It is possible to determine if a solution is acidic by

the red colour formed on the litmus when it is inserted into solution.

Better: An acidic solution turns blue litmus red.

(h) Textbook wording and phrases may be avoided to discourage the rote learner.

Poor : For every action there is equal and opposite reaction.

Better: A ball when struck against a wall rebounds due to Newton's third law of motion.

(i) Avoid the use of negative statements and especially double negatives to make a true statement false or a false statement true.

Poor : The Earth is not a planet.

Better: The Earth is a planet.

Poor : None of the articles in the Constitution is unnecessary.

Better: All the articles in the Constitution are necessary.

(j) Avoid items expressing opinions (unless you use T.F.O. variety).

Poor : The First Indian will go into the outer space next December. (opinion)

Poor : Planets and animals can live on Mars. (opinion)

(k) Avoid giving two ideas in one statement unless it depicts a cause-effect relationship.

Poor : Green leaves can manufacture food but colourless leaves cannot.

Better: Green leaves can manufacture food because they contain chlorophyll.

(l) Length of the true-false statements should approximately be the same.

(m) Number of true-false statements should approximately be the same.
However, some studies indicate that the number of false statements should be more (about 60%) than true statements because of their better discriminating power which may be due to the 'Acquiscence response set' i.e. a student is more likely to accept a statement than reject it when he is not sure of the knowledge involved in it.

1.4. *Steps in Constructing Objective-based True-False Items*

Superficially, it looks quite easy to frame True-False items. In fact, for developing a good item of this type some pre-requisites are necessary.
These are:

(a) Selection of a good proposition/concept/principle
(b) Restatement of the essence of proposition/concept/idea
(c) Finding implications of the proposition/concept
(d) Couching the antithesis of proposition in words and phrases
(e) Writing the item using the same essential point for one true and one false version.

Following example would clarify the underlying idea for framing such questions.

Concept: Structure of leaf is related to its functions.

	True	*False*
(a)	*Restate the essential idea in different words.* (Recalls)	
	(i) Plants floating in water have abundant stomata in their leaves.	(i) Plants floating in water are characterised by the absence of stomata in their leaves.
(b)	*Restate a part of original idea.* (Recalls)	
	(ii) Water plants with floating leaves have more	(ii) There are more stomata on the lower surface

stomata on the upper surface than on the lower surface.	than on the upper surface of the floating leaves of water plants.

(c) *Relate the basic idea to some other ideas* (Sees relationship)

(iii) Abundant stomata on upper surface of floating leaves are meant to enhance transpiration.	(iii) Abundant stomata on the upper surface of floating leaves are meant to reduce transpiration.

(d) *Develop implications of basic ideas* (Sees relationship)

(iv) Rate of transpiration in floating leaves is directly proportional to the number of stomata on the upper side.	(iv) Rate of transpiration in floating leaves is inversely proportional to the number of stomata on the upper side.

(e) *Infer the effect of different circumstances* (Infers)

(v) If all the stomata on the upper surface of the floating leaves of a water plant are smeared with vaseline, the leaves would rot.	(v) If all the stomata on the lower surface of the floating leaves of a water plant are smeared with vaseline, the leaves would rot.

1.5. *Usability of True-False Items*

The first criticism that is levelled against true-false items is their suspectibility to chance error resulting from guessing. A student with a blind guess would expect to get a score of 50%. Of course the assumption of blind guessing is not realistic because all the students are not likely to guess blindly all the items in a test. Secondly; such items are generally judged as testing for trivials. Thirdly; such items if culled out from the textbook encourage sheer verbal memory. Fourthly; many statements do not provide for enough background information or qualifications to judge an item as completely true or completely false. Fifthly; these items do not provide for explicit alternatives thereby making it difficult to judge the relative truth or falsity of the item. Sixthly; true-false items are considered psychologically harmful because they have a negative suggestion effect. Finally,

these items favour the aggressive individuals willing to take chance.

Inspite of the various shortcomings listed above, these items do have a number of advantages especially for classroom testing. Firstly; they provide for a simple, direct and fundamental test of students' knowledge. Secondly; true-false items are quite efficient because the number of scoreable responses per hour of testing time tends to be much higher than that of the multiple choice items. Thirdly; most teachers find the task of writing true-false items to be simpler and less time consuming as compared to equally good multiple choice items. Fourthly; they are less time consuming in construction due to less technical competence required. Fifthly; such items are good for testing cause-effect relationship and for discriminating between facts and opinions etc. (2)

1.6. *Improving Discrimination of True-False Items*

The most serious defect with true-false items is their amenability to the correct answer by chance without adequate understanding of the concept involved. When such a test is taken by students, the difference between those who attempt it correctly and those who answer it incorrectly is not as great as it should be. The reason is that those who attempt it correctly include students:

(a) who understand thoroughly the concept involved and deserve full credit for their answer;
(b) who do not have full grasp of the concept but are more or less lucky to choose the correct answer and therefore, do not deserve full credit as those in the first category.

Among those who answered incorrectly are those

(a) who were misinformed and therefore whose incorrect answer was expected;
(b) who had partial information relevant to the concept involved and their incorrect answer was more or less due to their bad luck.

To improve the discrimination of these items Ebel (1966) suggested five choices in place of two (T.F.) depending upon the extent to which an examinee may consider it true or false in accordance with the confidence he has about the correctness of his response. He may be given a single credit point with no penalty for items he chooses to answer for one mark. For double credit items he may be awarded 2 marks for every correct answer with a risk of double penalty for every incorrect answer to all such questions he chooses to answer with double credit. To discourage blind guessing he is allowed to omit an item without losing the score he could expect by blind guessing. Thus the credit or penalty he gets depends on his confidence about the correct answer. Thus responses are weighted as follows: (3)

Response Number	Confidence level of response	Score value		
		Right	Wrong	Omit
1.	The statement is probably true	2	—2	
2.	The statement is possibly true	1	0	
3.	I have no idea	—	—	0.5
4.	The statement is possibly false	1	0	
5.	The statement is probably false	2	—2	

The underlying assumption is that a student should not ask for double credit on an item unless he is confident enough that his answer would be correct in at least 2 out of 3 cases. So long as the proportion is above two thirds the students' score will remain higher if he chooses higher confidence level (double credit with double penalty) as illustrated below:

Example

Low (Possible) Score		High (Probable) Score
with 80%	$80 \times 1 = 80$	$(80 \times 2) - (20 \times 2) = 120$
with 70%	$70 \times 1 = 70$	$(70 \times 2) - (30 \times 2) = 80$
with 67%	$67 \times 1 = 67$	$(67 \times 2) - (33 \times 2) = 68$
with 66%	$66 \times 1 = 66$	$(66 \times 2) - (34 \times 2) = 64$
with 60%	$60 \times 1 = 60$	$(60 \times 2) - (40 \times 2) = 40$
with 50%	$50 \times 1 = 50$	$(50 \times 2) - (50 \times 2) = 00$

Thus a student should not opt for double credit on an item unless the odds are 2 to 1 or better.

1.7. *Score and Confidence Level*

Personality trait (confidence) has therefore some relationship with the score accrued on such a procedure. Bold, cautious and timid examinees may get different scores due to the difference in their personality trait (confidence). Who among these has advantage over the other is illustrated below:

(a) In a 40 item true-false test suppose a student knows enough about the quality of answers he can give, about 20 items, and his probability is $p=0.75$ (3 out of 4) but for the remaining 20 items, he is not that sure and his probability of attempting these items correctly is $p=0.60$ (3 out of 5). However, he is quite sure that he cannot answer these 20 items beyond $p=0.60$ and therefore, should not attempt at a probability more than $p=0.60$. So he attempts 20 items at $p=0.75$ and the other 20 at $p=0.60$.

(b) Student—A is overconfident and attempts all the 40 items with high confidence ($p=0.75$).

(c) Student—B is timidly cautious and attempts all the 40 items with low confidence ($p=0.60$).

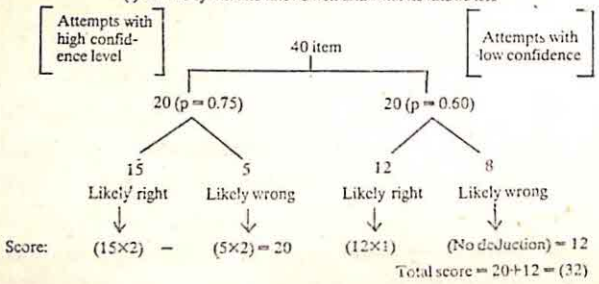
(d) Student—C is neither over-different nor over-cautious but unaware of what he knows well (High confidence) and what he knows poorly (Low confidence) and therefore attempts 20 items of $p=0.75$ with high confidence and 20 items of $p=0.60$ with low confidence using the weighted scores i.e. double credit (2 marks) with double penalty (−2) and single credit (one mark) with no penalty for items attempted at a high and low confidence level respectively. We may calculate the marks of the four types of students (in respect of personality trait i.e. confidence) to judge who among them would gain most.

From the scores of these three students it is clear that a student is likely to score more if he has

(a) knowledge of the subject, and

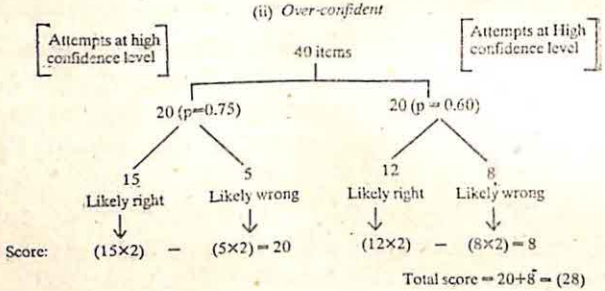
STUDENT – S

(i) *Aware of what he knows well and what he knows less*



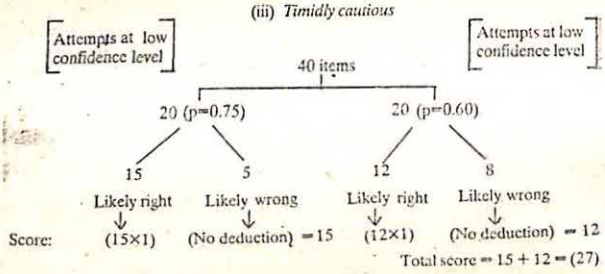
STUDENT – A

(ii) *Over-confident*



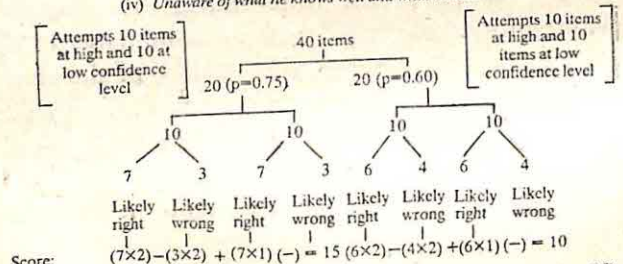
STUDENT – B

(iii) *Timidly cautious*



STUDENT – C

(iv) *Unaware of what he knows well and what he knows less*



(b) knowledge about the quality of his answers (confidence level).

His score will suffer when he is deficient in either. Distribution of students' score indicates whether a student is over-cautious, over-confident or unaware. Swineford (1941) derived a score to indicate the tendency of students to gamble. Thus score was derived by dividing the number of errors on items on which the examinee asked for double credit, at the risk of double penalty, by the total number of errors. Inter-correlations of these gamble scores on four tests ranged from 0.2 to 0.8, thereby revealing the tendency of students to gamble. (4)

1.8. *Ratio of True-False Statements*

Some investigators have found that false statements are more discriminating that true statements which may be due to the fact that when students are not sure of the correct response they tend to accept, rather than doubt, the truth or falsity of the statement. If this is correct, advantage may be taken to include more false statements (about 60%) than true statements. If more than that, say about 75% false students are included, then students are likely to recognize this imbalance while attempting and may take advantage of marking a statement false when in doubt. With weighted scoring, the reliability of true-false tests appears to increase without increasing the number of items as in conventional tests. Chronbach (5) and Thorndike have shown that reliability increased from 0.574, .765, .728 to .713, 0.828, .821 through this method. This reliability coefficient of 0.80 or higher for classroom tests of educational achievement is quite respectable. This method may be tried out to see how far it discourages the students' tendency to gamble through guessing.

Constant alternatives more commonly known as true-false items, are no doubt notorious for their chance scores due to 50% guessing possibility but none-the-less, these items are one of the most usable forms which, if constructed with care, to test objectives other than mere reproduction of facts can be used more profitably by classroom teachers. The weighted marks system discussed here can go a long way in improving

the discriminating power of true-false items. The utility of such items lies more in ensuring high-quality items during construction than in taking care of the guessing aspect by using correction formulae. However, increasing the number of choices from two (True-False) to four or five would definitely reduce the guessing possibility as in the case of multiple-choice items which are discussed in the next chapter.

2. Matching-Type Items

2.1. *Design*

Matching type of items are usually presented as a set of definitions, terms, events, names, phrases, etc., called the *stimuli* which are placed generally on the left hand side column, say I, and another set of names, terms, pictures, statements called the responses placed on the right hand side column, say II. Responses are to be matched with the stimuli. The items in column I are called *premises* and those in column II are called *responses*. Thus there are two lists and students are required to match each item with the corresponding response from the other list. Since each *item* from one list is to be matched with the corresponding *response*, it is considered as a one-test item which is counted as one score point. However, there are a number of items of similar nature which are to be put together so that the student is forced to make pairs out of the given responses.

The same kind of relationship should hold for each of the premises and the responses in the exercise. These relationships may be between a term and its definition, objects and its function, event and its date, author and its work, inventor and invention, cause and effect, problem and solution, examples and principles, etc. A matching-type item is in fact a multiple-choice item in which the items (premises) are different but the set of responses remains the same, out of which one response is to be selected unlike ordinary in multiple-choice items in which every item is followed by a separate set of 4 or 5 responses of which one is correct (key) in contrast to the matching-type in which every response acts as key for one or the other item in the list. Therefore, basically it is a modification of the multiple-choice

item. If the number of responses equals the number of items (premises) than it is termed as perfect matching and when the number of responses are more than the premises then it is called imperfect matching. In either case, the matching is to be done by the examinee on some basis or criterion which may be implicit or explicit. If there is a single criterion for matching the items with the responses then it is called *simple matching* but if more than one criterion is made the basis for matching vis-a-vis more than two lists given, then it is called *Compound matching*. The following illustrations would clarify the concept of different varieties of matching-type items.

2.2. *Varieties of Matching-type Items*

(i) *Column matching* (Imperfect Homogeneous variety)
For each term in column I select the statement from column II which best defines it and write its serial number in the bracket provided on the left against each item.

Column I	*Column II*
(*E*) 1. Curricular validity	A. Estimating correlation between parallel forms of the same test.
(*D*) 2. Construct validity	B. Estimating correlation between two measures obtained at the same time.
(*B*) 3. Concurrent validity	C. Estimating correlation between some measure and criterion measure obtained at a later date.
(*C*) 4. Predictive validity	D. Estimating how well the intended abilities hypothesised are actually measured.
	E. Estimating how well the content and instructional objectives are sampled and measured.
	F. Estimating relationship between two measures of the same person.

(ii) *Column matching* (Perfect Heterogeneous variety)
Match the discoveries in science given in column I with the names of disco given in column II and write the corresponding number in the bracket provided on the left against each of the items.

	I	II
(*B*)	1. Demonstrated circulation of blood	A. Pasteur
(*C*)	2. Demonstrated statistical approach to human heredity	B. Harvey
(*D*)	3. Demonstrated crucial experiments on the mechanism of heredity	G. Galton
(*A*)	4. Demonstrated experiment in support of spontaneous generation	D. Mendel

(iii) *Compound matching* (more than one basis—3 lists)
Three lists are given in three different columns. Select the matching animal and the class corresponding to each characteristic given in the column I and write their serial number in the same order against each item in the brackets provided to the left. (6)

	I	II	III
	Characteristics	*Animals*	*Class*
(F) (a)	1. Body covered with scales	A. Frog	a. Reptilia
		B. Pigeon	b. Mammilia
(C) (e)	2. Respires by gill	C. Fish	c. Aves
(A) (d)	3. Lives both on land and in water	D. Butterfly	d. Amphibia
		E. Bat	e. Pisces
(E) (b)	4. Suckles its young ones	F. Snake	f. Arthropods

(iv) *Classification variety* (Re-usable responses)
Identify whether the sentence is simple, compound or complex and write against each item in the bracket provided, on the left.

A. if the sentence is simple
B. if the sentence is compound
C. if the sentence is complex

(*A*) 1. Sidhu went to his school on bicycle.
(*C*) 2. If Anjum were in Delhi she could have visited the Asiad with her friends.
(*B*) 3. In Delhi summer days are long and nights are short.
(*A*) 4. Navdeep is an obedient boy.
(*B*) 5. During the 9th Asiad new stadia were errected and new flyovers built.

(v) *Alternative response matching*

In this case each response may be relevant or irrelevant to the corresponding item and the examinee is required to indicate by letter or serial number whether or not the response corresponds to the specified item.

Example

In the blank before the number of each datum write the letter R if the hypothesis is relevant to the datum and write I if the hypothesis does not support the datum or is irrelevant.

	Datum	*Hypothesis*
I.	1. With above-average students the correlation between scores on intelligence test and creativity test is 0.00	A. Creativity is related to intelligence.
R.	2. Scores of 30 year-olds tend to be slightly higher than the 15 year-olds	B. Some mental abilities tend to increase after the age of 20.
R.	3. I.Q. scores of identical twins raised apart tend to correlate higher than the I.Q. scores of fraternal twins raised in the same family.	C. Intelligence is an inherited trait.

(vi) *Check-list variety*

It requires the student to go through a list of phrases, statements or words and tick mark the ones that correspond (match) to the criterion or the base which is indicated in the direction of the items.

Example

Examine the list given below and tick-mark the words or phrases which relate to estimation of the reliability of a teacher-made test.

1. Stability	.....	7. Equivalence	.....
2. Concurrent	.....	8. Split half	.....
3. Predictive	.....	9. Stability and equivalence	.....
4. K.R.—20	.....	10. Judgemental	.....
5. K.R.—21	.....	11. Construct	.....
6. Facility index	.....	12. Discrimination index	.....

(vii) *Master-list* (*Single Key*)

It requires the students to choose from a constant group of alternatives, usually 3 to 5, the one that applies best to a group of items. The key response may be reused or one of the responses may not act as a key for any of the items in the group. Thus a master-list resembles the common two-column matching items in having the same set of responses reusable for different items (Premises). It differs from it in that the choice of options remains the same (constant) for each item unlike the traditional matching-type items in which after attempting each item the number of options (responses) does not remain constant but goes on becoming less and less after each item. So choices can be used more than once for different items. Moreover, the number of items is much more than the number of keys (master-list). These items differ from multiple choice items in that the same set of choices (key list) are usable, unlike in the multiple choice items in which alternatives change from item to item.

Example

From the list (master key) at the top, select the kind of plants which show marphological features corresponding to each of the items that follow:

A. Desert plants B. Alpine plants C. Water plants

D. Marshy plants

(*B*) 1. A well-developed root-system with reduced leaves on stem.
(*A*) 2. A deep root-system, succulent stems with reduced leaves.
(*D*) 3. Presence of respiratory roots and capacity to withstand salt water.
(*B*) 4. Roots spread near the surface with stems having needle-like leaves.
(*C*) 5. Poorly-developed roots with abundance of air cavities.
(*A*) 6. Flattened green stems with leaves modified into spines.
(*C*) 7. Dissected roots with broad leaves having stomata on upper surface.

(viii) *Master-list* (Combined Key)

Example

Some words are italicised in each of the sentences given below. Put in bracket.

A. if the word is used as a noun.
B. if the word is used as an adjective.
C. if the word is used both as a noun and as an adjective.
D. if the word is used neither as a noun nor as an adjective.

Sentences

(*D*) 1. Slow and steady *wins* the race.
(*C*) 2. The *Race* course in Delhi is being utilised for horse races.
(*A*) 3. *Velodrome* is a stadium for cycle races.
(*B*) 4. *Thirty three* nations took part in the ninth Asiad held at Delhi.
(*C*) 5. Women *hockey* players played *hockey* well against Japan.
(*A*) 6. Indian women gymnasts gave a good *performance.*

2.3. *Usability*

Compactness of matching items makes it a difficult form of

questions in terms of space and testing time. These are good for making a rapid survey of certain subject matters like those of leading personalities, group of events, definitions of basic terms etc. Such items are good for testing students' ability to identify objects presented graphically or pictorially, translate terms and definitions from one form and establish relationship between two sets of objects, concepts, events, etc. A major criticism of these items is that they can be used to measure only lower level learning outcomes like recall of dates, events, etc. But this is not true, as the above examples should convince us that the relationship between problems and solution, data and hypothesis, characteristics and classification, cause and effect, etc. can be tested provided that the item writer is well trained to develop objective based items.

Nevertheless, if sufficient care is not taken, matching items may encourage a serial-memory action rather than association and relationships. It is also difficult sometimes to find a homogenous cluster of items which in turn need homogeneous responses. Some common errors that are found in such items are: (a) vague direction, (b) vaguely stated premises, (c) usually long statements or responses, (d) lack of homogeniety in responses, and (e) difficulty of students regarding mode of responding or keying the items. If the following points are kept in mind while constructing the items, the quality of such items can be greatly improved.

2.4. *Steps in Construction*

The following steps are possible in constructing matching exercises:

(i) Decide what is the intended aspect of measurement.
(ii) Make a list of the premises to be used.
(iii) Identify the responses including the distractors.
(iv) Arrange the items in some order.
(v) Write the direction for pupils.
(vi) Indicate the space and mode of responding.

2.5. *Suggestions for Constructing Matching Items*

The following suggestions may be kept in mind while construct-

ing items:

(i) Be cognizant of the objective and content elements which form the table of specification of the proposed test.
(ii) Indicate clearly the basis of matching in the direction of the items or in the headings of the columns.
(iii) Have one of the lists consisting of short phrases, single words, or numbers for quick examination and reducing the reading load.
(iv) Use only homogenous and closely related materials in any one matching exercise.
(v) Use heterogenous stimuli and responses if the item is to be set at an easy level.
(vi) Make the number of alternatives larger (about 1½ times) than the number of items to reduce the chance of guessing for the items attempted towards the end.
(vii) Keep each list short, ranging from 5 to 12 premises and responses, the optimum size being 5-8 items per exercise.
(viii) Avoid clues of any kind.
(ix) Avoid grammatical inconsistencies like 'all men', 'all women', all plural, all singular, etc. to minimise relevant clues.
(x) Every premise must have one clear plausible answer in the other column.
(xi) Label each set of premises and responses, especially when a separate direction is not given to help the examinee to understand his task well.
(xii) Give clear instructions if some of the responses can be used more than once.
(xiii) Avoid framing long sets of items for matching exercises lest they should overemphasise an objective on which it is based.
(xiv) Arrange response chronologically, alphabetically or logically to facilitate pupils finding the response.
(xv) Keep the matching set on the same page to avoid confusion in matching items with responses given on two pages.

3. Rearrangement Type

3.1. *Design*

Such items require the pupils to rearrange the randomly presented material into some specified order. The material may be presented in the form of a series of statements one after the other or in the form of responses as given in the multiple choice type, i.e. A, B, C, D etc. A direction is provided as to whether the responses are to be rearranged by writing them in the specified order, or to serial them into a particular order or to indicate the serial number of each response etc, etc.

3.2. *Types*

(i) *Chronological order*

Example
Rewrite the serial number of the events given according to chronological order in the space provided.

1. V.V. Giri
2. Zakir Hussain
3. Sanjiva Reddy
4. Radhakrishnan

Rewritten as: *3 1 4 2*

(ii) *Functional order*

Example
Rearrange the following steps involved in the manufacture of food by plants, in order of their occurrence and write the serial number in the bracket provided against each.

1. Formation of PGAL (5)
2. Formation of A.T.P. (2)
3. Excitation of chlorophyll (1)
4. Splitting of water (3)
5. Evolution of oxygen (4)
6. Formation of starch (6)

(iii) *Logical order*

In developing a blue print of a unit test, the following steps are adopted for preparing a unit test. Rearrange the serial number of each step in the blank provided against each to indicate the correct order.

S. No.	*Steps*	*Correct order*
1.	Writing the questions	4
2.	Deciding about the total marks for the test	1
3.	Identify major concepts to be tested	2
4.	Editing the items	6
5.	Reviewing the test	7
6.	Formulate unit objectives relating to concepts identified	3
7.	Developing the model answers	5

(iv) *Developmental order*

In certain cases, the content matter is more amenable to a developmental order which may be natural, evolutionary or of any other form. Students are required to reorder the words, statements, steps, etc., according to the specified criteria.

Example

The evolutionary developmental order of the following animals is given in order from 1 to 5. Rewrite the sequence showing correct order of development.

1. Bat 2. Amoeba 3. Star fish 4. Pigeon 5. Shark fish

(Incorrect)

1. Amoeba 2. Star fish 3. Shark fish 4. Pigeon 5. Bat

(Correct)

(v) *Hierarchical order*

In terms of increasing complexity the facts, concepts, principles, etc., can be arranged to depict a hierarchical order. Students may be asked to reorder the jumbled or randomly-placed responses in order of hierarchy.

Example

Rearrange the given order of educational objectives of (i) Bloom, (ii) Krathwhol, and (iii) Simpson in terms of hierarchical learning, by indicating the serial number against each objective listed.

(i)		(ii)		(iii)	
1. Knowledge	(1)	1. Organisation	(*4*)	1. Set	(2)
2. Application	(3)	2. Characterisation	(5)	2. Perception	(1)
3. Analysis	(4)	3. Valuing	(*3*)	3. Complex overt response	(5)
4. Evaluation	(6)	4. Receiving	(*1*)	4. Originating	(6)
5. Comprehension	(2)	5. Responding	(*2*)	5. Mechanism	(4)
6. Synthesis	(5)			6. Guided response	(3)

(vi) *Taxonomic order*

In this case the responses are arranged in order of certain categories or classes based on a natural relationship as we know in case of classification of plants and animals.

Example

Examine the following order of classification of certain animals and rewrite the correct order in terms of natural relationships (i) or indicate the correct order by rearranging the serial numbers (ii).

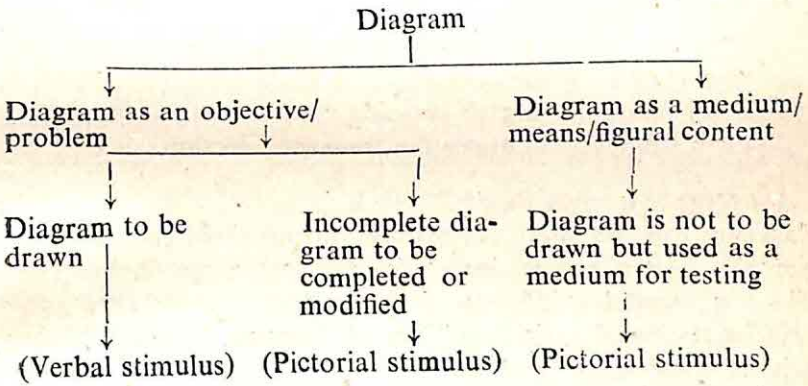
(i) (Jumbled)	(Rearranged)	(ii) (Jumbled)	(Rearranged)	
1. Phylum	*Kingdom*	1. Animalia	Animalia	*1*
2. Kingdom	*Phylum*	2. Primates	Primates	*3*
3. Order	*Class*	3. Hominidae	Hominidae	*4*
4. Class	Order	4. Mammalia	Mammalia	*2*
5. Species	*Genus*	5. Sapiens	Sapiens	*6*
6. Genus	*Species*	6. Homo	Homo	*5*

4. Pictorial Varieties

4.1. *Design*

Students are presented with pictorial or graphic stimuli on which items are based. A picture, chart, table, sketch, diagram or a graph can be given as a stimulus followed by responses in case of selection items. In case of free response questions, the questions can be put direct on the given pictorial stimulus. Thus objective-type, short-answer type and even essay-type questions can be set on the chosen picture or diagram which acts as a base or medium for testing various abilities. Likewise a picture of a diagram can be used as a problem by itself in which the solution lies in the drawing of a diagram, in its completion or modification depending upon the purpose of the exercise. Thus the two types of diagrammatic questions or items can be classified as under:

4.2. *Classification*



4.3. *Varieties of Pictorial Questions/Items*

A. *Diagram as objective*

In this case diagram is to be drawn on, completed, modified, rectified, developed or labelled. Objective is to test students' skill in drawing.

1. *Diagrammatic sketches* (magnified or reduced in size)
   (i) Draw an electron microscope diagram of an animal cell.

2. *Faithful diagrams* (actuals)
   (i) Draw a diagram of the experimental set-up placed before you on the table.
   (ii) Draw a diagram of the material as you see under the microscope. (a slide is given for identification)
3. *Incomplete diagrams* (missing parts)
   (i) In the given diagram, showing the longitudinal section of a bisexual typical flower, draw the missing structures.
4. *Modified diagrams* (rectifying)
   (i) Four sketches representing various stages showing a mitotic division are given. Identify the wrongly drawn/wrongly labelled parts and re-draw/re-label the same correctly.
5. *Representational diagram* (data-providing)
   (i) Draw a diagram showing the population trend in India given below:

Year:	1947	1950	1960	1970	1980
Population in millions:	400	450	500	550	650

6. *Outline diagrams*
   (i) Draw an outline political map of India and fill up the capital cities of each state.
   (ii) Draw an outline sketch showing the internal structure of a dicot stem as seen under the transverse section.

B. *Diagram as a means/basis*

Diagrams can be used for testing various skills and abilities. Examples for testing drawing skills have been given above. The following examples illustrate how diagrams can be used as a basis for testing of various abilities. A diagram is provided in each of these cases and the pupils are required to observe or examine the diagram/s and answer the question, based on the diagram/s.

4.4. *Ability tested*

1. *Recall*

   In the given diagram label the parts 1, 2, 3 and 4.

2. *Recognise*

   In the given diagram of a plant cell, label cell wall mitochondrium and chloroplast.

3. *Translate*

In the labelled floral diagram given below which part bears the same relationship to stamens as label-2 to the carpel?

4. *Cite example*

The proximal food constituents are represented in the form of a diagram given opposite. This diagram is an example of a

A. flow chart.
*B. pie graph. (Pie graph provided)
C. pictograph.
D. table.

5. *Identify relationship*

Which organelle in the given diagram of a fresh water amoeba is concerned with osmoregulation?

6. *Compare*

Which of the following two diagrams labelled as A and B is that of an animal cell? (Two diagrams are provided).

7. *Classify*

Which of the leaves (4 diagrams of leaf-tips showing reticulate venation in 3 and parallel in one are given) does not belong to the same group to which the others belong?

8. *Detect errors*

Find the error if any, in the experimental set-up shown in the diagram for preparation of oxygen gas.

9. *Interpret*

What does this graph indicate regarding the rate of photosynthesis? (A graph is provided)

10. *Explain*

Observe the experimental set up for the preparation of oxygen gas in the laboratory (A diagram is given). Why do we not keep the delivery tude-A dipped in the solution like that of thistle funnel-B?

11. *Read instruments*

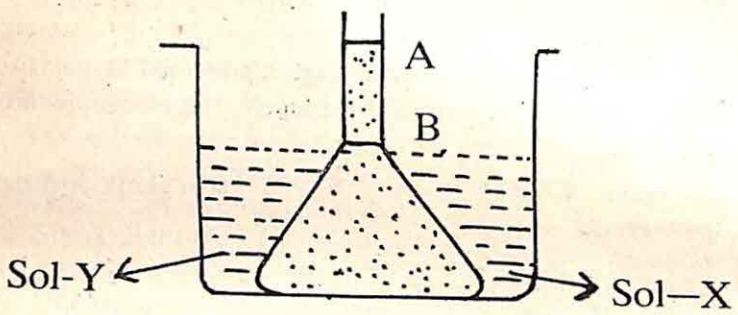
Observe the readings indicated in the three thermometers. In which case it is the highest and in which case it is the lowest?

(Diagrams are provided)

12. *Analyse*

In the given diagram, the level of solution-Y comes down

from A to B after 3 hours. This could be possible only if we assume that the concentration of solution-X is



*A. more than solution-Y
B. less than solution-Y
C. equal to solution-Y
D. nothing to do with the fall in level.

13. *Hypothesise*

In the diagram given in item 12 above, if the solutions X and Y are interchanged the level in thistle funnel rises from B to A. When more of water is added to solution-X a stage comes when there is neither rise nor fall in the level. From these observations what hypothesis would you frame to explain these observations?

14. *Suggest procedures*

Identify the hypothesis underlying in the experiment shown above and suggest an alternative experiment to demonstrate that hypothesis.

15. *Reason*

In the 3 sets of diagrams X, Y, and Z showing the effect on the shape and turgidity of a freshly-cut potato slice is indicated. How would you account for this change?

Fig. X
Empty beaker

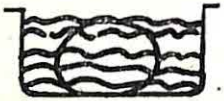


Fig. Y
Beaker having water

Fig. Z
Beaker having sugar solution

16. *Infer/conclude*

Observe the two stages shown in diagrams A and B of plant cell given below. What inference you can draw about the effect on (Plasmolysis) cell content and the physiological process involved in it.

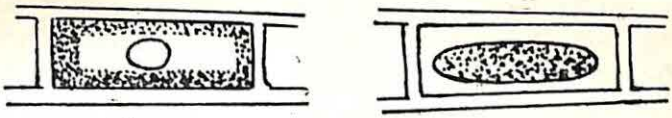
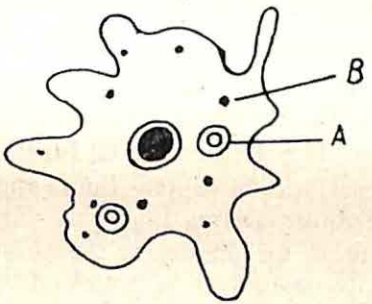


Fig A Fig B

17. *Predict*

In the diagram given below if this species is transferred to the sea and gets adapted, which of the following observations is most likely to be true?



Fresh water amoeba

A. Organelle A and B would both disappear.
*B. Organelle A would disappear and no effect on —B.
C. Organelle B would disappear and no effect on —A.
D. Neither organelle—A nor organelle—B would disappear.

5. Structured Questions

5.1. *Concept*

The nomenclature of structured questions is rather confusing. Different people use it differently and as such their connotation

also varies from one context to the other. In different texts they are referred to as 'thought questions', 'broken up essays', (8) 'split essay', 'connected set of questions', 'subsequential sub-questions', 'passage-based questions' etc. These questions are infact as old as the reading-comprehension testing whereby a few questions based on a given para used to be set to test the reading-comprehension of students. Thus they are not a novel variety but only a rediscovered or re-emphasised variety that has gained more attention of the evaluators to guage the wide-ranging abilities of students. With the increased emphasis on the development of higher-level abilities in our instructional programme, the need for greater use of structured questions in the assessment of pupils' achievement is being increasingly realised. It is, therefore, useful to get an insight into the form and functions of structured questions.

5.2. *Format*

A typical structured question is one which is rooted in a given or stipulated situation providing the needed premises, material, or data that provide the uniform content to be used as the medium of testing. The basic material forms the introductory statement that prescribes the content limits and direction to the examinees. The examiner desires that this introductory information must be used for answering the questions that follow. Thus the introductory statement is followed by a number of sub-questions based on the subject matter forming the basis. An initial question is set, followed by a few more sub-questions in a sequential order. Each sub-question relates to a specific ability desired to be tested through the given item that remains the same for every sub-question. The diagram (Fig. 1) given on the next page represents the format of structured questions.

5.3. *Characteristics*

The diagrammatic sketch given on the next page, may help us to derive the following characteristic features of structured questions:

1. The same introductory material is used for every sub-question.
2. Introductory material may be presented in the form of

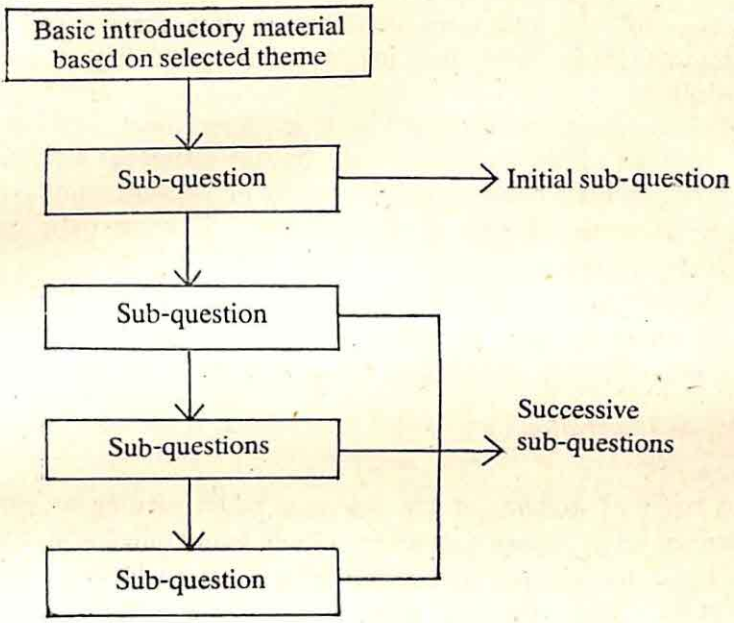


Fig. 1: Format of the Structured Questions

a statement, a passage, table, diagram or experimental data.

3. Each sub-question is based on a complex problematic situation and not a usual textual situation.
4. Introductory material should neither lack any material without which an answer is not possible, nor should it provide extra information which is not useable and adds to the reading load of students.
5. The subject-matter content is based on one specific theme.
6. Sub-questions may be structured in the form of the short answer, restricted (broken) essay or objective type.
7. Usually the same form of sub-questions is used in one set of questions linked to the central theme.
8. Each sub-question must demand reading of introductory material for answering i.e. one should not be able to answer a question without referring to the introductory statement.
9. Each sub-question is independent of the other and is not dependent on a knowledge of the previous question.
10. Each sub-question tests a different ability as far as possible.

11. None of the questions attempts to test mere recall of factual information but involves testing of higher level abilities.
12. Sub-questions are arranged in a predetermined particular order or criterion which may be the estimated difficulty level, abilities being tested, hierarchy of concepts involved or in terms of form of sub-questions, if more than one form of questions, is used.
13. Depending upon the abilities to be tested, a suitable form of question may be used. However, objective type and short-answer type questions are more useable and profitable.

5.3.1. *An illustration*

Passage

A batch of students of life sciences, while visiting a game sanctuary in a reserve forest saw a buck being killed and eaten by a tiger. Earlier they had seen a herd of bucks feeding upon the grass. They also saw a peacock catching a snake. Under the tall trees a leaf-less, golden yellow vine was found closely attached to the twigs of a green shrub. This vine, named dodder (Cuscuta) showed some spots caused by a fungus.*

Read the above passage and answer the following questions:

21. In the above ecosystem if the grass is the producer the secondary consumer is
    1. buck.
    2. tiger.
    3. peacock.
    4. snake.
22. The relationship of the fungus with Cuscuta is that of
    1. hyperparasitism.
    2. parasitism.
    3. mutualism.
    4. cannibalism.
23. The inter-relationship of grass, buck and tiger as revealed in the above passage is termed as
    1. flow of energy,
    2. food pyramid.

*Source: An old N.T.S. question paper.

3. food web.
4. food chain.

5.4. *Useability of Structured Questions*

'The use of structured questions depends on the purpose of the test, the nature of examinees, subject-matter to be tested and abilities to be tested. (9) We may take each factor one by one.

Purpose of the test determines to some extent the nature of the questions. Achievement tests aim at measuring the status of the individual in terms of instructional objectives. A diagnostic test aims at discovering inadequacies in learning and their causes. A predictor test, like an aptitude test, aims at collecting such evidences about students' that help them to predict the likely future success of individuals as in tests used for selection purposes e.g. National Talent Search Examination of the NCERT. Structured questions can be used effectively in all the three types. However, in certain cases e.g. in a practical test of science the use of structured questions is not profitable. Since our discussion is focussed on achievement testing in which the language and content of a subject is assumed as a learning experience, the use of such questions is very much limited for diagnostic tests and much less in predictor tests.

The nature of the group determines the homogeneity or heterogeneity of the students. In a heterogeneous group, question testing different abilities or questions at various levels of achievement are needed for which structured questions are very useful. The smaller groups are likely to be more homogeneous and therefore require questions at almost the same level. Thus, when large heterogeneous groups are to be tested, these questions are very useful. Likewise, when several groups, which are similar, are to be tested at different times for comparison over the years in the same institution, structured questions can be fruitfully used.

When compared to the essay-type, the greater objectivity in the marking scheme is another favourable point when marking is to be done by a large number of examiners as in the case of public examinations. As structured questions demand some

minimum level of reading comprehension and the writing ability of a pupil, the use of such questions should be done carefully, keeping in view the students' ability to comprehend the written material at a reasonable speed.

The subject matter also affects the construction of such questions. If the question is heavily loaded with subject matter vis-a-vis the concepts being tested, then most of the questions set on that material would be deeply embedded in that content. They can, however, be made independent of the content to a great extent. In other words, the minimum relevant content is provided for equalisation of content, thereby allowing every candidate to answer, using the same content. Thus different abilities may be tested through these questions.

Of central importance is the abilities that should be measured through structured questions such as the pupils' ability to understand, apply, analyse, interpret hypothesise and evaluate, besides testing their creative abilities. Hudson (1973) gave the following example (Fig. 2) in which we find an introductory statement followed by sub-questions. Each question is set on the higher level intellectual abilities, of course assuming the basic knowledge and concepts regarded as pre-requisite to the answering of questions. Each question is an independent measure and is not dependent on any other question, thereby measuring independent abilities (10).

Structured Question – Biology

The production of heat as a by-product of metabolic processes may be demonstrated using apparatus as shown in the diagram on the next page.

Soaked live pea seeds are placed in one flask and an equal quantity of soaked peas which have been killed by steeping in mercury (II) chloride solution are placed in the other. The rise in temperature caused by metabolic activity of the live peas causes an expansion of the air in that side of the apparatus and thus a pressure difference which is recorded by the water manometre tube.

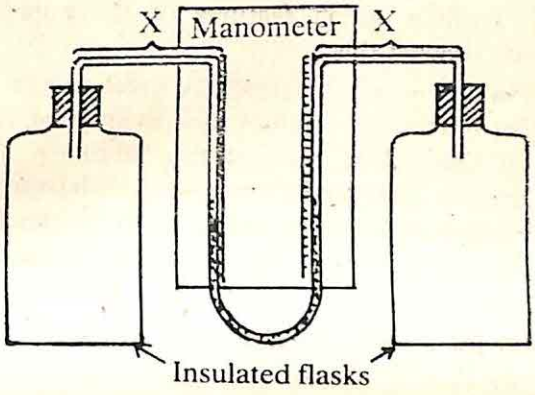


Fig. 2
Source: *Assessment Techniques—An Introduction*, B. Hudson, pp. 100-107.

Several different people setting up this experiment proceeded in different ways.

(i) One placed equal numbers of peas in each flask, another equal weights and a third equal volumes. What are the disadvantages of each of the three procedures?

(ii) One person used metal insulated flasks while another used clear glass insulated flasks. What differences would you expect in the results and why?

(iii) One person set up the apparatus with the tubes X each approximately 100 mm in length while another used tubes over 500 mm in length. Which person would be likely to obtain the more accurate results and why?

(iv) One person placed only a small quantity of peas in each flask while another almost filled each flask with peas. Which factors should be considered in this experiment in determining the most appropriate quantity to be used?

(v) Even given ideally matched quantities of pea seed and the most appropriate form of flasks and tubing there are factors other than thermal expansion which could affect the volume of air on each side. Describe one such factor and its probable effect.

(vi) This experiment can be used as a qualitative demonstration but could not be used in any quantitative investi-

gation. Explain which features of the apparatus and procedure prevent this.

It is observable that all the questions, given in the example, test the higher intellectual abilities such as analysis, interpretation, reasoning, cause-effect relationship, inference, prediction etc. All the six questions lend themselves to relatively objective marking and each question is independent of the other, thereby making assessment more reliable. From this illustration, it is quite obvious that structured questions can be made use of quite profitably for testing varied intellectual abilities.

When to use such questions and when not to use would therefore depend on the purpose of the test, the nature of the subject-matter, the nature of the group and the abilities to be tested.

5.5. *Presentation*

When the series of structured questions are to be constructed, one can take advantage of the type of stimulus presented and the possible reception sense like sight, hearing, touch/taste and smell. Stimulus can be presented in the form of words, apparatus material, tables, graphs, numbers, symbols, pictures, diagrams etc. If all questions are presented in words, of course with sight as the reception sense, then the pupils might feel bored. In case of long tests, it is advisable to present the material in a varied manner to cater to pupils' interest. This would also ensure a more accurate measure of students' abilities. Comprehension of material in the least possible time with the minimum effort, is what is required of a good introductory statement. The material may, therefore, be presented through a diagram. The introductory statement must be phrased in simple, comprehensible language. This ensures more reliability by attributing the performance of students to their developed abilities rather than being contaminated by their language ability to comprehend. Real, authentic and accurate descriptions should be utilised, rather than fictitious data for which a variety of source material may be investigated. Use of a familiar diagram, stereotyped and difficult language would spoil the

quality of the structured questions. Concise, comprehensible and accurate content presented in an appropriate form are the pre-requisites to a good introductory statement.

6. Marking Schemes

As in any other type of questions, it is always advisable to write the answer at the same time as the preparation of the question. In fact, it is better if a complete answer is written as is expected from the examinee. This helps to obviate various types of flaws and inadequacies in the answers and also helps in chiselling out the question itself to arrive at the intended answer. Although complete objectivity of marking requires that all possible correct answer should be pre-determined and written in the marking scheme, it is, however, not any easy task, especially when open-ended questions are used, which allow freedom of response to some extent. If the sub-questions are properly structured, keeping in view the ability to be tested and are pinpointed using definite language, it is possible to write a precise and objective marking scheme. It is not possible in case of the essay-type and such short-answer type questions that allow more freedom of response on the part of examinees.

If alternative answers are possible to the questions, then the question itself should be re-examined and revised. Likewise, if the credit points in a particular answer are very many, then that sub-question may be reviewed to make it more pinpointed or it may be split up into two or more separate sub-questions. A good structured question may require not more than one, two or three marks for each sub-question.

The marking scheme may be a personal marking scheme when it is to be used by the individual teacher for students in his own class or it may be a marking scheme which may be used by more than one examiner as in the case of a public examination. In both the cases, it has to be worked out very carefully and special care may be taken in the case of the personal marking scheme because it will not be discussed with other colleagues for its finalisation to make possible amendments.

While developing the marking scheme, care may be taken that each value point may be given proportionate weightage

out of the total allocated marks, say 10. The total for these marks should invariably lead to exact marks allocated for that question.

There are some questions where objectivity in marking is not really possible and are in fact marked entirely on a subjective basis. It is only in structured questions that the writer can frame questions and develop the marking scheme in a manner that minimises subjectivity in marking. In spite of this, the examiner may find an anwer which may be correct according to him although that answer may not be indicated as the intended answer in the marking scheme. In such cases, marks may be allotted by the examiner in terms of the intended answer.

Head (1967) showed that in such cases the rank order of the pupils taking a particular test is hardly affected even though the range of total marks awarded may differ significantly when the same answers are marked strictly according to a scheme which is inflexible.(11) Marking is sometimes a subjective process, but it should not be based entirely on personal judgement. A good marking scheme should however, indicate criteria of judgment alongwith allocation of marks for each criterion as clearly as possible.

Once the marking scheme is written, one may like to forget it for some time and keep it away. Since the first draft of a question is far from being perfect, the marking scheme can be reviewed after a few days. It is better if reviewing and editing of questions is done by more than one independent persons who are well conversant with the item-writing technology besides possessing mastery of the subject.

7. Limitations

In spite of the fact that structured questions find a prominent place in today's literature and are even used in selection tests, they have some inherent limitations. These questions are very difficult to construct for two reasons. First, it is really a tough job to develop on introductory passage which could really provide the needed base material and is at the same time not a

simply replication of textual material. Secondly, all the sub-questions related to the main should also not only test the higher level abilities but should also not overlap and interlock. It is thus a difficult task to frame such questions.

These questions demand a lot of time in structuring—good questions that serve the intended objective and at the same time are economical from the point of view of testing.

They involve a lot of reading load on the part of the examinees and, therefore, quite a good deal of time is wasted in reading and re-reading the passage in order to answer the questions that are set on the passage. For every question the examinee has to go through the passage, again and again.

Besides, these questions are not economical in the sense that for testing one major theme or central issue, one has to utilise three to five questions. Moreover, when questions are used as a part of the annual public examination, one is forced to restrict the content to only one unit and therefore, proportionate emphasis on different topics may not be ensured. As such the content-validity of a test in which too many structured questions are ued, is likely to be reduced.

Since these question involve a lot of students' ability to comprehend, it presupposes a certain level of comprehension on the part of students. If they have not developed a minimum level of comprehension such questions cannot be used. Therefore, their use at lower stages, like the primary stage, becomes impossible.

8. To Sum Up

From the foregoing discussion on structured questions, it can be inferred that these questions can be used only for certain purposes for pupils of a particular maturity level. These questions are very useful when some complex problem or a major theme is to be thoroughly investigated and demands the interplay of various higher level intellectual processes that provide evidence about the student's command of certain specific concepts and principles.

Their use at the unit-testing level cannot be over-emphasised because of their focus on testing different abilities ranging from

the simple to the complex. However, their use in selection tests can be questioned on the basis of the time-requirement per item. Structured questions presently in vogue, if analysed, would reveal that more that 50% of the sub-questions can be answered, irrespective of the basic introductory material or the passage provided. This indicates the inherent difficulty of framing independent test-item testing different abilities based on the given passage. Their use in classroom testing is quite diagnostic to discover students' inabilities. Moreover, their use in the form of oral questions based on a particular diagram or data which can be presented to the class on the blackboard can be of much use. Thus, structured questions are not economical to use unless one attempts to test the higher level abilities. Nevertheless, these questions provide a good challenge to the item writers as well as to the teachers who can make use of these questions effectively after having acquired the technical knowhow of their construction as well as the usability of such questions.

CHAPTER XII

QUALITIES OF A GOOD MEASURING INSTRUMENT

Whenever we measure students' achievement by using a teacher-made test or a standardised test we use the data for making decisions on the assumption that the measuring instrument is of good quality. In the case of standardised tests, the manual of such tests gives the coefficient of validity and that of reliability but in most of the teacher-made tests, the results of examinations are used without taking into cognizance the quality of the instrument used. In fact most class teachers are not conversant with the characteristic of a good examination. Once they appreciate the need for understanding the characteristics of a good achievement test, they would be able to interpret test results more scientifically. After going through this chapter the reader would be able to:

(a) define usability, reliability and validity of measuring instruments in his own words,
(b) identify relationship between reliability and validity with respect to the nature and scope of each and factors affecting them,
(c) classify various sources of variance relating to individual, administration and chance factors,
(d) estimate reliability coefficients using different methods and understand their limitations in usability of each method, for interpretation of test results,

(e) define, describe and classify various types of validity and the implications of each for test construction, and
(f) apply the concept of reliability and validity in judging the quality of the measuring instruments.

The three major requirements of a good test in this chapter are validity, reliability and usability. Each of these is discussed in this chapter. Since validity is comparatively a more comprehensive concept than reliability and reliability is more important concept than usability I would prefer to start discussion of these concepts in the reverse order.

1. Usability

A test is said to be usable if it is esay to construct, easy to administer and easy to interpret. An essay test is easier to construct and administer than a test of multiple-choice items but it is more difficult to interpret than a test consisting of multiple-choice items. Administrability of evaluation refers to the ease and accuracy with which the evaluators and the examinees can follow the directions. The requirement of directions vary from test to test. Morevever, if the test is too time-consuming to administer, like the Stanford-Binet-Intelligence Test, its usability is restricted. This test is administered individually and requires about 1 hour for the examinee and half an hour for the examiner. The directions are so complex and detailed that it requires one semester course to learn to administer and score this test. Essay-type tests are easy to administer but administration of the objective test requires more paper vis-a-vis expenditure and has to be administered separately with a separate time limit. Therefore, it has to be distributed and collected separately with a number of instructions to the examinees. Tests involving sub-tests requiring separate time limit for each sub-test are more susceptible to errors in timing and require the administrator of the test to have a time-piece during administration.

Some tests require no supervision. Oral instructions may vary from room to room if these are not given in the test itself or are read from the paper. When given in the test itself, they may be separated from the test material so as to avoid distrac-

tion. It is better to fit the length of the test to the classroom period. In case of the standardised test, the cost of the test is another factor besides the format of the test which may facilitate the mode of responding to the test questions and/or writing the responses at a specified place designated for the responses. Tick-marking or encircling a true-false, item against each statement is easier if the blank space or the word T and F are provided on the right instead of to the left of each statement.

Using a separate answer sheet in case of the objective type-test enhances the usability of the test as it becomes easier to analyse the result besides possibility of reusing the test material. The cost of the analysability of the scores, the complexity of the operation, the provision of a scheme for marking and summation of responses, do reflect on the administerability of the test.

Interpretability refers to the degree to which the scores of a test are readily derived and understood. Applying a scoring key to get raw scores, listing of credit or value points in the marking scheme, use of correct formulae, machine or manual scoring, arrangement of items in the test in terms of objectives, content, form of questions, difficulty level (estimated) etc. have a bearing on the interpretability of scores. To give meaning to the raw score, the use of the group-norm or the age-norm would help in interpretation. The use of a referent i.e. self, group or criterion is another variable that will determine the mode of interpretation. Should we estimate or predict? Should we compare with the norm to look for deviation of students' score from the norm or should we compare with the predetermined criteria of expected performance in terms of mastery of learning? Are we interested in grading students, providing remedial instruction or judging instructional efficiency. These are some of the cogent factors that make the scores meaningful for interpretation purpose.

Thus usability involves judging of distractions, distributions, of papers, carrying on instructions, timing, supervision, ending of the test, making notations, security, malpractices, etc. for administration of a test. The order of scoring, rescoring, marking schemes, scoring devices, keys, machine and hand scoring, recording, tabulating, summarising, and profiles are relevant to

the scorability of the test. Precautions like rapport between the examiner and the examinees, preparation of examinees, obtaining material administration of tests to pupils, scoring of tests and the use of test results are the different phases of a testing programme which have to direct or indirect bearing on the usability of the test.

2. Reliability

The more important universal criterion of a good measuring device is the reliability. Whereas validity deals with what to measure, reliability is concerned with how efficiently does it measure. The former deals with relevance of the purpose, objective or the intended outcome, while the latter concerns itself with the accuracy and consistency of measurement. Reliability is an essential part of validity. Both relevance and reliability constitute validity. Therefore, validity to a certain extent depends on the reliability of the test but the vice-versa is not true. A test may be totally invalid, yet reliable because of its consistency and accuracy of measurement. A test must be reliable in order to be valid.

Whenever any measurement is made, it has to be repeated once more or a number of times to be sure that our measurement is accurate and precise. But it is seldom that we find, on repitition, the measurement to be the same. There is always some variation, although the extent of variation would vary depending upon the precision of the instrument, the user of the instrument and many other factors.

I poured three bottles of petrol into my scooter five times in a month but the mileage was not always the same. It was 32 Km, 31 Km, 32 Km, 33 Km, and 32 Km, per litre. I kept a record for a month. As the traffic and the weather conditions were the same throughout the month there should have been no change in the mileage. Had there been a constant mileage of 32 Km, I would have been satisfied that it was the true mileage. But why did all the five measures vary, although the true mileage (32 Km) did not vary? This may have been due to chance errors which might have crept into due to one or more of the following reasons:

(a) Three litres of petrol may not be exactly three litres indicated at the petrol pump; it may be a little more or a little less.
(b) Over or under-estimation of the amount due to reading of the indicator that shows only 1/10th of the division of each litre.
(c) The milometer might not have registered accurately the distance travelled.
(d) The reading of the meter may not have been accurate every time, sometimes the nearest digit was rounded off to the next and sometimes to the last digit.
(e) Difference due to observation of the level of the tank, by the attendant etc. etc.

There may be many more such factors. Still we find that the mileage averaged was fairly constant. Such a consistency is desirable in our educational measurement. A good measuring instrument should yield dependable information or scores when used repeatedly in unchanging situations.

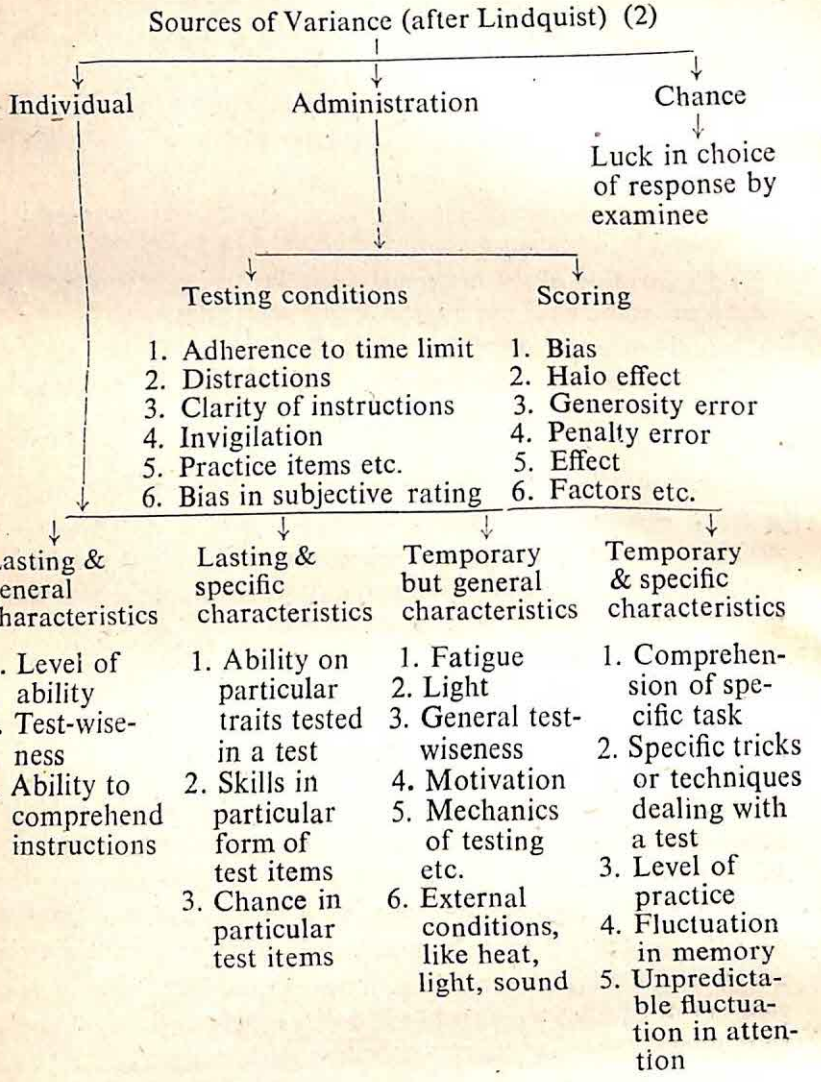
2.1. *Sources of Variance*

Scores on a test do not remain constant when the same or similar tests are repeated. This is due to systematic or unsystematic variations caused by systematic or unsystematic factors. Systematic variations are characterised by some trend, orderly pattern, regular increase or decrease in scores, age, sex, training, growth, etc. On the other hand factor like fatigue, seating arrangements, disturbance or noise, poor flow of the pen etc. are random in character and cause changes which vary from moment to moment, thus causing unsystematic variations.

The various sources of variance categorised are given on the next page.

2.2. *Definitions*

1. By reliability we mean the consistency of the measurement of a particular achievement from time to time or from test to test (Ebel) (1)



2. Reliability..... .....is simply the extent of unsystematic variation in the quantitative descriptions of the amount of some trait an individual possesses or manifests when that is measured a number of times (Ghiselli) (3)
3. Reliability of any set of measurements is logically defined as the proportion of their variances, that is, true variance (Guilford) (4)

4. Reliability concerns the precision of measurement regardless of *what* is measured. Some random error is involved in all scientific measurements..... Psychological measures also contain a portion of random error analogous to that of the mental ruler (Nunally) (5)
5. "Reliability concerns the accuracy and consistency with which it measures whatever it does measure in the group with which it is used. To be valid—that is, to serve its purpose adequately, a test must measure something with reasonably high reliability, and that something must be fairly closely related to the function it is used to measure (Thorndike) (6)

2.3. *Why Scores Needs to be Known?*

Scores need to be known for comparison among individuals—assignment of individuals to groups—prediction of other types of behaviour from one score—comparison of traits and abilities of an individual as in guidance or assessment of effects of systematic factors on an individual's performance, i.e. for dependable deductions concerning test properties. Usual test practices based on the assumption that all examinees are predictable.

2.4. *What is Reliability?*

A test is for obtaining a score for an individual. Hence what to measure is assumed. The next question is, how to measure? How will it measure *accuracy* of measurement or *precision* of measurement? The accuracy with which a score represents the status of an individual in whatever aspect the test measures him.

The ideal test tells the truth consistently. Hence the score should remain the same over repeated measurement, or if another similar test is administered, the same score should be obtained.

Reliability and validity are different but interdependent concepts. Validity pertains to what a test measures; reliability pertains to how it measures. Validity pertains to relevance; reliability refers to accuracy. Validity is known by agreements with an *external* criterion; reliability is known by an *internal* criterion. *Consistency* or stability are other words which describe reliability.

2.5. *Why Do Scores Change?*

Scores do not remain constant over repeated administrations or for similar tests. Scores change due to *systematic factors* and *unsystematic factors*. Systematic factors cause *systematic variation* characterised by an orderly progression or pattern, with the scores obtained by an individual changing from one occasion to the next in some trend. Systematic changes appear as a regular increase or decrease in scores, or they may follow some cycle. Sex, age, or place of birth are some systematic factors. Learning, training, growth are some other factors which show systematic variation.

Unsystematic factors like environmental, acquired or native determinees of behaviour produce unsystematic variations, which are random in character. The changes may be from moment to moment. Some examples are fluctuations in attention, inconsistent or a bulky pen, lighting, seating, ill health, etc.

Reliability of measurement is the extent of unsystematic variation in the quantitative descriptions of the amount of some trait an individual possesses or manifests when that trait is measured a number of times.

If in a series of *parallel tests*, the standard deviation of scores for a single individual is zero, then the reliability of measurement is perfect. If this is small, reliability is small. This is the *standard error of measurement*. It is the standard deviation of scores of one individual over a series of parallel tests. If the SEM is 3, which means for 68 per cent cases, the scores may vary by 3 points. Thus this index gives an indication of the absolute accuracy of test scores.

Another way to describe it is to say how well scores for a group of individuals remain consistent with two parallel tests. This is called the *reliability coefficient*. It is the index of the extent to which scores on any one parallel test can predict scores on any other. Operationally, it is a kind of self-correlation of a test.

2.6. *True Scores and Error Scores*

Unsystematic factors contribute to the scores we get, in some degree or other. Hence the score we get for an individual can be described as:

Obtained score = True score + Error Score
(which may be positive or negative)

This may be expressed as:

$$X_t = X_\infty + X_e$$

where:

X_t = obtained score
X_∞ = true score
X_e = error component

Three assumptions underlying the above are:

(a) The individual possesses stable characteristics or traits that persist through time.
(b) Errors are completely random, i.e., error scores are uncorrelated with scores on any other variable, and with true scores.
(c) Obtained scores are the result of the addition of true and error scores.

"The existence of error scores leading to error variance, corresponds to the fact of unreliability, and its amount relative to the total of all variance is a measure of the degree of unreliability." (Lindquist)

If σ^2_∞ is the variance of the true score, σ^2_e is the variance of error of measurement and magnitude of error of measurement is unrelated to the magnitude of true score then:

$$\begin{matrix} \sigma^2_t & = & \sigma^2_\infty & + & \sigma^2_e \\ \begin{pmatrix}\text{(obtained}\\ \text{variance}\end{pmatrix} & & \begin{pmatrix}\text{variance of}\\ \text{true scores}\end{pmatrix} & & \begin{pmatrix}\text{variance of}\\ \text{error scores}\end{pmatrix} \end{matrix} \qquad \text{(i)}$$

These variance can be related to reliability coefficient by means of the following equation:

$$\text{Reliability Coefficient } r_{tt} = \frac{\sigma^2_\infty}{\sigma^2_t} \qquad \text{(ii)}$$

Therefore, reliability is the proportion of true variance in

obtained test score.

Since $\sigma^2_{\infty} = \sigma^2_t - \sigma^2_e$ we can substitute the difference in this to get $r_{tt} = 1 - \frac{\sigma^2_e}{\sigma^2_t}$

Since σ_{∞}^2 and σ_t^2 are not known, other formulas have been devised, for estimating true variance.

The relationship between standard error of measurement and reliability coefficient is expressed by the equation

Standard error of measurement or
(or standard error of obtained score)

$$SEM = \sqrt{1-r} \quad OR \quad \sigma_e = \sigma_t \sqrt{1-r_{tt}}$$

Reliability Coefficient from Item Statistics

r_{tt} depends upon how the split halves have been obtained. Kuder-Richardson formulas depend upon item statistics. Each item is to be considered as a separate test. Hence all items are to be obtained of equal difficulty and inter-correlation. The formula is

$$r_{tt} = \left(\frac{n}{n-1}\right)\left(\frac{\sigma^2_t - \Sigma pq}{\sigma^2_t}\right) \qquad (7)$$

where n = number of items in the test.

p = proportion of correct responses to each item

q = 1 − p

The reliability is obtained on a single administration of test.

Reliability from Analysis of Variance

In this, as in the above method, arbitrary methods of choosing the particular two halves are dispensed with. All the information about consistency of performance from item to item within the test is made use of to obtain an estimate of internal consistency (See Lindquist, p. 590). The analysis of variance approach is more useful for obtaining reliability estimates from items or trials scored with a range of scores.

2.7. *Factors Affecting Reliability*

2.7.1. *Factors associated with the test itself*

(i) *Item homogeneity*

If the items are homogeneous with the test, higher intercorrelations, and therefore higher reliability, is ensured. A perfectly homogeneous test will have the highest reliability but the validity of the test will suffer. K-R formulas depend on the complete homogeneity of function from item to item.

(ii) *Item difficulty*

The reliability of a test is the highest if the items are of median difficulty, or have a difficulty level of 0.5 (in Ebel, Guilford). When the items vary widely, a smaller reliability coefficient is obtained. Hence in short tests, items of the difficulty level 0.5 can be pooled. In longer tests, some spread of item difficulty is desirable. It is better to have a rectangular distribution as is achieved in such a test (Guilford). In the case of multiple-choice items where chance makes it easier to get higher scores, the difficulty level of items may be greater than 0.5, say 0.6 (Guilford).

(iii) *Type of items*

Higher reliability will be obtained in the case of multiple-choice type items.

(iv) *Items based on memory*

If items are based on *memory*, reliability will not be high as compared to items testing higher objectives.

(v) *Length of the test*

The reliability of the test increases with increase in the number of items, or the length of the test. The items added should resemble the old ones, in form and content, and should have an equal difficulty level and equal intercorrelation with one another and with the old items.

(This amounts to combining two parallel forms) (See Guilford, p. 352). Both the K.R. formula and the Spearman-Brown Prophesy formula suggests why this happens.

From Spearman-Brown formula

$$r_n = \frac{nr_s}{(n-1)r_{s+1}} \qquad (8)$$

"r_n is the reliability of test in items as long as a shorter test of known reliability r_s is equal to n times the reliability of the shorter test, divided by (n−1) times the reliability of the shorter test plus 1."

It is possible to calculate the number of times a test has to be lengthened to get the required reliability coefficient. (See Ebel, p. 315). If the initial reliability is lower, the first increase in length will result in a larger gain than after successive increases in length. If the length has to be increased by many times the original size to get optimum reliability, other methods have to be adopted.

In the case of multiple-choice items, when the number of response options is increased, the reliability of the test also increases as shown by Remmers and others (see Lindquist). Thorndike points out that a multiple-choice item behaves as if it were composed of k—1 two-choice items where k varies from 2 to 5.

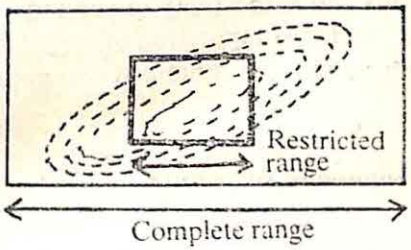
In the case of subjectively scored items, such as essay-type questions, increasing the number of raters or examiners has the same effect as that of increasing the length of the test.

In many tests, the length may have to be decreased to increase the reliability. Discarding poor items will achieve this purpose. The formula is useful in another way. It enables us to see if we can safely shorten a test without appreciably lowering the reliability. (Guilford, F.S., p. 466).

2.7.2. *Factors associated with the group*

Range of ability

Test reliability depends upon the range of ability of the population sample taken for administration of the test. If the sample is heterogeneous, the reliability coefficient will be higher. If the range of scores is narrower, the reliability coefficient will be smaller. This is depicted in the following diagram given by Guilford. (9)



Complete range. Correlation for the cases within the smaller rectangle is much smaller than that for cases within the larger rectangle

Hence reliability should be found out from administration over a sample of high range of scores or greater variability.

Reliabilities of tests reported in manuals have to be evaluated against this background. If the test had been administered to children from one particular grade, the reliability reported is generally on the lower side, whereas reliability got from administration to a single age-group is on the higher side. (See Thorndike and Hagen, p. 186) (See Guilford, p. 392 for a formula for estimating the reliability of a test when used in a population of one SD from knowledge of its reliability in a population with a different SD).

A test if very difficult or very easy will not have a high reliability. If difficult, it leads to guessing, and hence low reliability. If easy, very low variability, and hence poor reliability. The test may have different reliabilities at different score levels. If the sample is too small, there will be less variability, and hence the reliability will be low.

2.7.3. *Factors associated with test administration*

(a) Provision of instructions in the test
(b) Provision of practice exercises
(c) Accuracy of timing
(d) Sufficiency of time
(e) Standardization of motivation
(f) Conditions of test taking

(g) Time between trials
(h) Scoring methods and subjectivity of scoring
(i) Method employed for computation of reliability coefficient by which split-halves are obtained or parallel forms are got.

2.8. *Methods of Estimating Reliability*

At least two sets of measurements have to be obtained for the purpose. The correlation coefficient is obtained between the two sets of scores so obtained.

(i) *Test-retest method or repeated testing method*

There are certain definite advantage in this method. Only one set of sample tasks or items is needed. This holds constant in both the applications and hence no variance attributable to changed task is introduced. The two applications may be done with some interval or without any interval of time. If there is no interval of time, the carry-over effect from one testing occasion to another will be present (the memory factor). If the interval is large, it is eliminated but changes in the individual take place. The problem of getting the same group is great. The attitude of the examinees also may be a new factor that comes into play.

Hence, in measures of intellect, temperament, or achievement, where a large number of tasks or items forming the universe are used, the test-retest method is not defensible. (10)

Test-retest is usually higher than that based upon parallel forms.

(ii) *Parallel forms or equivalent forms*

This method gives the coefficient of equivalence and is considered a preferred procedure for estimating reliability. It goes well with the theory of the true and error score and the *eclectic* concept of true scores and parallel tests (see Ghiselli, p. 281).

What are parallel test forms? Each form should be built to the specifications, and consist of random samples of items. It should have same distribution of item difficulties and same distribution of item-test correlations. The Thorndike equivalent

forms are tests having identical true variance and no overlap of error variance. (See Ghiselli for distinction between parallel forms and parallel tests, p. 279).

If parallel forms are administered one after the other without interval, the reliability resembles the split-half reliability or internal consistency reliability. If they are administered with a long interval, the reliability resembles test-retest reliability.

The advantage is that the error due to sampling of particular items in the test is taken care of. There is no specific carry-over effect due to memory. The disadvantage is the change in content—a source of error variance. The difficulty of building more than one form is another disadvantage. The parallel-form reliability is usually an under-estimation of reliability due to this factor. This method is particularly suitable for speed tests. (See Ghiselli, p. 222).

(iii) *Subdivision of single, total test*

It gives the internal consistency reliability of the test. There are different methods of obtaining two-half lengths: (a) two-halves equivalent in content and difficulty, (b) alternate items leading to two halves by splitting is the most frequently used method when the items are arranged in order of difficulty, (c) alternate groups of items leading to two halves, (d) the first half forming one part and the second half forming the second part, provided the test is not heterogeneous in content. Whichever method is used the parts are to be parallel, i.e. (equal average difficulty items, equal spread of difficulty, equal item intercorrelation. The same amount of time should be devoted to each part).

The correlation coefficient obtained in such cases is reliability for half the test length. Spearman-Brown's Prophesy formula (See Lindquist, p. 580) (or any one of the other formulae) is applied to get the reliability for the whole test. The formula is based on assumptions about equality of SDs of part scores and the quality of the part score inter-correlations. This method is economical in terms of effort and time. However the disadvantage is its applicability to power tests only. (See Lindquist, p. 582). Lack of time. interval between taking the two halves (they are infact intermingled) gives variance

which is mostly systematic. Successive items are quite alike also. It gives a misleadingly high reliability coefficient. This reliability is a function of the choice of the split.

Sources of variation represented in different procedures for estimating reliability*

Sources of variation	Immediate retest, same test	Retest after interval same test	Parallel test form without time interval	Parallel test form with time interval	Odd even halves of single test	K.R. analysis single test
How much the score can be expected to fluctuate owing to:						
(a) Variations arising within the measurement procedure itself	x	x	x	x	x	x
(b) Changes in the individual from day to day	—	x	—	x	—	—
(c) Changes in the specific sample of tasks	—	—	x	x	x	x
(d) Changes in the individuals speed of work	x	x	x	x	—	—

**Source*: "Measurement and Evaluation in Psychology and Education" by Thorndike and Hagen (1969), p. 186.

2.9. *Interpretation of Test Reliability*

Standard Error of Measurement

Since the reliability coefficient is dependent upon the variability of the group and not merely on the quality of the test, the *standard error of measurement* is used sometimes in place of the reliability coefficient. It can be used even when a group differs from the original group.

"Standard Error of Measurement (SEM) is equal to the Standard deviation S_x of the test times the square root of 1 minus the reliability of the test." (11)

SEM	=	S_x	$\sqrt{1-r_{xx}}$
Standard error of measure-ment		Standard deviation of the test	Reliability coefficient

It provides an indication of the absolute accuracy of the test scores.

Interpreting: If the standard error of measurement is 3, then for 68 per cent cases, the scores will be 3 points more or less.

How high should reliability be?

For evaluating reliability coefficient, the following data are needed:

(a) the particular set of operations by which consistency of performance was obtained,
(b) the description of the group tested,
(c) statistical characteristics of the group, and
(d) adequacy of sampling.

There is no hard-and-fast rule as to how high reliability should be. It can be low for research purposes (Guilford,

p. 389). More crucial is the validity of the test. For practical purposes of diagnosis and prediction it should be higher, even at the cost of validity. It is possible to have a reliability coefficient of 0.90 for good objective-type tests but 0.80 is considered a standard (Ebel, p. 330).

3. Validity

3.1. *Some Definitions*

(i) *Lindquist* (12)
"The validity of a test may be defined as the accuracy with which it measures that which it is intended to measure, or the degree to which it approaches infallibility in measuring what it purports to measure." According to this definition, to determine how valid a test is, one must compare the reality of what it DOES measure with some ideal conception of what it OUGHT to measure.

(ii) *Cureton* (13)
"The validity of a test is an estimate of the correlation between the raw test scores and the 'true' (i.e. perfectly reliable) criterion scores". This definition does not answer the question, "What should the test measure", Criterion scores that measures exactly the same thing as the test is intended to measure are seldom available, even in an unreliable form for most of the classroom tests of educational achievement.

(iii) *Edgerton* (1949)
"By validity we refer to the extent to which the measuring device is useful for a given purpose."

(iv) *Chronbach* (1960)
"The more fully and confidently a test can be interpreted, the greater its validity."

(v) *Guilford* (1946)
"In a very general sense, a test is valid for anything with which it correlates."

(vi) *Websters New Collegiate Dictionary* (1951, p. 940)
"A test or anything else, is valid if it is 'Founded' on truth or fact; capable of being verified, supported or defended; well grounded, sound."

(vii) *Gulliksen* (1950)
"The validity of a test is the correlation of the test with some criterion."

(viii) *English and English* (14)
"A property of the whole measuring or testing process but specially of the testing instrument that ensures that the obtained test scores correctly measures the variables they are supposed to measure. Validity is always validity for measurement of a particular variable. There is no such thing as general validity, nor is there any absolute validity—we determine the degree of validity and the validity index has no meaning apart from the particular operations by which it is determined."

From the above mentioned definitions, we can infer that evaluators differ in their concept of validity. Some specialists specify correlations with a criterion as Gulliksen. Some require estimation of corrected coefficient as Cureton. Others emphasise accuracy in relation to user's intent. Still others relate it to utility and interpretability of test scores. It, therefore, seems difficult to arrive at a common conceptual definition and even if one is formulated it may be so abstract that no useful purpose would be served. In fact these is need for operational definition of the term validity.

3.2. *Types of Validity*

(i) *Concurrent validity*
Validity is concerned with the relation of test scores to an accepted contemporary criterion of performance on the variable which the test is intended to measure—e.g. correlation of raw score on test with teachers' ranking or on any standardized tests.

(ii) *Construct validity*
Validi ty is concerned with "what psychological qualities a test

measures" and is evaluated "by demonstrating that certain explanatory constructs account to some degree for performance on the tes.t"

(iii) *Content validity*
Validity is concerned with the adequacy of sampling of a specified universe of content.

(iv) *Curricular validity*
Validity is determined by examining the content of the test itself and judging the degree to which it is a true measure of the important objectives of the course, or a truly representative sampling of the essential materials of instruction.

(v) *Empirical validity*
Validity refers to the relation between test scores and criterion scores, the latter being an independent and direct measure of that which the test is designed to predict.

(vi) *Face validity*
Validity refers, not to what a test necessarily measures, but appears to measure.

(vii) *Factorial validity*
The validity of a test is the correlation between that test and the factor common to a group of tests or other measures of behaviour. Such validity is based on a factor analysis.

(viii) *Intrinsic validity*
Validity arising from the fact that the items of a test inherently and necessarily call for the behaviour that the test as a whole is designed to measure.

(ix) *Sampling validity*
Is a measure of content validity obtained by determining how far the test items are a representative sample of the universe of behaviour that defines the variables to the measured.

(x) *Status validity*
It refers to the correspondence between test scores and the

concurrent status of tests.

(xi) *Predictive validity*
Validity is concerned with the relation of test scores to measures on a criterion based on performance at some later time.

(xii) *Validity by definition*
For some tests the objective is defined solely in terms of the population of questions from which the sample comprising the test is drawn e.g. when the ability to handle one hundred number facts of addition is tested by a sampling of those number facts.

Although all these validities are not distinct from each other, still some differences appear to justify their classification into two major categories.

3.3. *Validity Classified*

Ebel classifies all types of validities into two major categories viz. Direct and Derived.

(i) *Direct Validity*
A test has direct, primary validity to the extent the tasks included in it represent faithfully, and in due proportion, the kinds of tasks that provide an operational definition of the achievement or the trait in question.

(ii) *Derived Validity*
A test has a derived, secondary validity to the extent that the scores it yields correlate with criterion scores which possesses direct, primary validity.

Thorndike and Hagen (15) have dichotomised all these types into two; those which depend primarily on rational analysis and professional judgement (Direct validity); and those which depend on empirical and statistical evidence (Derived validity).

Direct Validities	*Derived Validities*
(a) Validity by definition	(a) Empirical validity
(b) Content validity	(b) Concurrent validity
(c) Curricular validity	(c) Predictive validity

(d) Intrinsic validity
(e) Face validity

(d) Factorial validity
(e) Construct validity

This division is not sharp but convenient for classification.

From the above definitions, we find a number of terms used as qualities of validity. The three main purposes for which psychological measurement is used are:

(a) Representation of a specified universe of content.
(b) Measure of psychological traits.
(c) Establishment of a functional relationship with a particular variable.

Accordingly we have three types of validity and can classify various terms under each type:

Validity

Curricular	*Construct*	*Predictive*
1. Intrinsic	1. Trait	1. Post diction
2. Content	2. Factorial	2. Concurrent
3. Rational	3. Convergent	3. Prediction
4. Face	4. Nomological	4. Empirical
5. Sample	5. Discriminant	5. Statistical
6. Representativeness		6. Practical
7. Logical		7. Classical
		8. Status

It is again obvious that there are some common conceptual elements in these definitions but at the same time it is only possible to have a loose and general definition because of striking differences in these definitions. What is actually needed is more concrete and realistic conception of the constituents that make a test qualitative.

Another paradox is that although the test specialists regard validity as the most important quality of a test, yet we find it lacking in most of our present tests according to the comments of the specialists. It can only be due to lack of understanding of the validity concept.

Likewise, another difficulty arises due to questioning the validity of a measuring procedure rather than looking for an

operational definition of quantitative concepts as in the case of physical measurements: An atom clock for example is not justified on the basis of its superior validity but because of its closer approximation to the measurement of true time that reflects, in fact, superior reliability. Therefore, instead of asking for the validity of the basic method of measurement, which is meaningless, we should be satisfied with operational definition of the thing to be measured.

3.4. *Can We Define Validity?*

Validity is regarded as the most important characteristic of a good measuring device. It refers to the degree to which a test measures what it purports to measure. This definition needs further elucidation: What does a test actually measure; to what extent, in what situations, under what conditions and to what do it do its job? In a Biology test, for example, a test is designed to measure an instructional outcome like the understanding of concepts. This is its validity but if it measures something else also, perhaps drawing skills, its validity to measure understanding of concepts is reduced. This test may be highly valid for 10th class students but comparatively less valid for 11th class students and still less for college students. Thus a test may be a highly valid measure for predicting pupils' academic success in class XI but less valid for predicting success for admission to a medical college and a still less valid measure of success for a medical on the job. Therefore, a test is valid for a certain purpose and may be wholly invalid for another purpose.

A test is never completely valid or entirely invalid. We are mainly concerned with how valid is the test how well it serves the purpose. It may do so extremely well, reasonably well or hardly well.

3.5. *Implications of Validity*

Before taking up the problems and difficulties in determining the validity of our tests, it may be useful to understand the implications of the three major types of validity.

3.5.1. *Curricular validity*

When a test is based on the content and instructional objectives of the course, its curricular validity depends on the degree to which the test items represent these aspects. Therefore, if on analysis of the test content and the analysis of the course content and instructional objectives, we find that the former represent the latter, it is the curricular validity of the test. Because analysis is based on rational judgement this type of validity is also called rational or logical validity. It is also called content validity or sample validity because this analysis is mainly in terms of course content. However, it may not be construed that content is the only thing we are interested in, in devising an instrument. It is both the content (concepts, learning points) and the process (skills, abilities) that represent the curricular and hence needed for curricular validity.

In recent years, there has been a growing interest to emphasise in instruction, the ability to apply, interpret and evaluate thereby de-emphasising verbalism. As such, in test construction too, it is necessary that the instrument must attempt to measure these abilities rather than factual information in order to have curricular validity. But this is possible only when the objectives of instruction go beyond the memorising of facts. When this is done, both the curricular validity of test and instruction would improve. We may summarise the curricular validity by its following features:

(a) Judgement is based on subjective standards and therefore, may be quoted as high, moderate or low.
(b) How far an item is relevant to the trait being measured and the extent to which the whole set of items measure all aspects of the trait. The former represents the first approximation of its validity and if it bears a high correlation with the total score, the item is worth retaining.
(c) Adequacy of the content and objectives has to be ensured through proper representation of each by proper designing of the test on the basis of experts' judgment and users of the test.

3.5.2. *Empirical validity* (concurrent and predictive)

This type of validity is based on the concept that a good test must yield scores that agree with some outside criterion such as teachers marks of future success on the job.

The correlation between the test scores and the criterion scores determines the validity. When the criterion measure is taken at the same time or nearly at the same time it is termed concurrent validity. For example, when teachers' rating or a standardised test is taken as a criterion against which pupils' score on a teacher-made test is judged, it represents concurrent validity or status validity. Correlation between test scores and some measure of performance or success obtained later is termed predictive validity. Here the criterion is future success on the job or performance measure at a later date. For example, when test scores in a high school biology test are correlated with pupils' performance later on at the medical college or test scores at the final M.B.B.S. examination in relation to success as a doctor, it gives the predictive validity of the test used.

Concurrent and predictive validity are primarily empirical or statistical evaluation. For such evaluation it is essential that the criterion should represent the same quality, attribute or ability as the test, being validated, is supposed to measure. Most of the intelligence, aptitude and achievement tests are validated in this way as a valid measure of some ability, quality or trait.

(a) *Limiting factors*

No doubt we have performance records, teachers' ratings, standardised tests and proficiency tests which may be used as criteria for validation; but how far are these criteria valid in themselves. If the test scores are reliable and teachers ratings are a reliable and valid measure of students' performance, then the correlation between the test scores and the teachers' ratings will be an accurate measure of the validity of teachers' ratings. Therefore, in order to have maximum empirical validity of a test, the criterion (teachers' rating here) must be a reliable and valid measure besides ensuring the reliability of the test itself.

(b) *Criteria of a criterion*

All criteria are partially valid because they measure only a part of the success on the job. A test on an Engineering course for selection of engineers gives only a particular criterion of success as an engineer. The ultimate criterion is one's success in life as an engineer. But this criterion of success in life, like other criteria, is inaccessible and as such we have to be satisfied with the substitute for it. One problem is to choose the most valid criterion from among the available ones. We may look for the following qualities in a criterion measure.

(i) *Relevance*

A criterion is relevant to the extent the score on the criterion measure is determined by the same factors that constitute success on the job. This is judged on rational and expert judgement.

(ii) *Freedom from bias*

A criterion should provide each person with the same opportunity to make a score. It means control of working conditions too: that criterion scores depend on the factors in the individual rather than factors in the conditions of work.

(iii) *Reliability*

This pertains to the stability or reproducibility of the measure of success. If the criterion performance itself is unstable then it cannot be predicted by any other test.

(iv) *Availability*

Depending upon the cost, the time factor and convenience, the criterion measure has to be selected for its practicability.

(c) *Interpretation and use of predictive validity*

The amount of validity is the degree of relationship between the predictor and criterion scores. The size of the correlation is a direct indication of the amount of validity. We cannot expect perfect or even high correlations in most cases. Only moderation correlation between the predictor(s) test(s) can reasonably be expected. A validity coefficient, should be interpreted and used in applied settings in terms of the extent to

which it indicates a possible improvement in the average quality of persons that would be obtained by employing the instrument in question.

A test having predictive validity can be used in situations like the following:

(i) Prediction of events occurring at a previous date on the basis of questionnaire administered to adults.
(ii) A test designed to predict success in college means forecasting but a test used to predict brain damage is not meant to forecast who will suffer brain damage in future; rather who does and who does not have brain damage at the time of test administration.
(iii) When a test is used to predict future performance, it is an aptitude or prognostic test.
(iv) Diagnostic tests indicate sources of maladjustment, pupils' difficulties in learning, showing thereby the existence of favourable state of affairs.

3.5.3 *Construct validity*

A construct is an attribute or human characteristic on the basis of which we assume that such an attribute accounts for some human behaviour. Constructs like dominance, insecurity, rigidity, inability to apply are used in psychology and help in identifying the psychological concepts they represent. Whenever an evaluation instrument is believed to reflect a particular construct, its construct validity must be found out. A construct must be well defined in order to draw an inference which can be verified. Construct validity is employed where curricular or empirical validity cannot be used or to supplement it. According to Chronbach, "construct validity is involved whenever a test is to be interpreted as a measure of some attribute which is not operationally defined". The problem is, what constructs account for variance in test performance? Construct validation is essential whenever no criterion or universe of content is accepted as entirely adequate to define the quality being measured.

Constructs are essential for the development and testing of theories in Science. Science is mainly concerned with developing

a measure of constructs and finding functional relationships between measures of different constructs. Construct very to the extent the domain of related observable variables are big or small and lightly or loosely defined. The larger the domain, the greater is the difficulty of defining it for allocating it to a particular domain. Measurement and validation of a construct involves three processes:

(a) to specify the domain of observables;
(b) to determine the extent to which all or some of those observables correlate with one another or are affected by experimental treatments; and
(c) to determine whether or not, some or all measures of such variables act as though they measure the same construct.

Relation between the observables

In an empirical investigation two approaches are followed to see how well the measures of observables go together. In the individual difference approach, scores are obtained on some of the measures. Each measure is correlated with every other measure and analysis of correlation is done to find out the extent to which different measures measure the same thing. In the controlled experiment approach, we see how far the treatment conditions have a similar effect on the different measures of observables. The degree to which two measures are affected similarly by different experimental treatments, they are assumed to be measuring the same thing as the construct.

From this when we find high correlations among the measures, it indicates that all measures measure the same thing. Measures forming clusters indicate that a number of different things are being measured and a zero correlation among measures indicates that different things are being measured by different measures. On the basis of evidence got from this, modifications can be made in the specification of the domain of observables of a construct besides adducing the relationship between the two constructs.

3.6. *Limitations in the Use of Validity*

Keeping in view the types of definitions of validity given by

various authors, it is necessary to develop some operational definition of validity. A more concrete and realistic conception is needed, of the qualities that make a test good. Ebel has raised the issue whether we are justified to question the validity of the method of measurement. What is needed according to him is the operational definition of quantitative concepts and the accuracy of measurement.

Another difficulty that arises is how to know "what the test is supposed to measure." Criterion measures are seldom devised painstakingly when compared to the labour put into the construction of the test.

Moreover, what is the criterion of the criterion measure used to determine the validity of criterion measure? Unless a test is supposed to measure an ideal quantity which is measurable directly and therefore defined operationally, there would be difficulty in validating it etc.

In predictive validity, we judge a test—say, T, as a measure of a trait I on the basis of the extent to which it predicts trait C when C is a function of T and V or many more variables, rather than only a function of T. Can we justify the quality of a barometer merely on the basis of its accuracy in forecasting the weather? Likewise, can we put our faith in the validation of a test which contains items to test understanding and ability to apply against teacher's subjective rating in the form of grades given on the basis of pupils performance in biology, arithmetic, physical education, craft etc? The flaw is obvious. The fallacy arises from the assumption that the relation of a measure to one, single, other measure (criterion) is all important. Unless it is related to other measures it is functionless. What is needed is the meaningfulness of the test which depends on operational definitions, of measurement procedures, reliability, usability and convenience of the test in use.

CHAPTER XIII

UNIT TESTING—A DEVELOPMENTAL APPROACH

National Policy of Education (1986) envisages a continuous comprehensive evaluation for making evaluation an integral part of instruction to provide regular feedback to students about their adequacies and inadequacies in learning. This can be ensured by using various devices among which a unit test is the most potential information gathering devices which a class teacher can use to measure students' learning and diagnose difficulties. But it requires technical competence on the part of teachers to develop such unit tests and use them effectively for various purposes.

It is, therefore expected that after going through this chapter a teacher should be able to:

(a) understand the concept of a learning unit and a unit test,
(b) identify various types of content elements from a learning unit,
(c) formulate specific objectives of the unit of teaching/ learning,
(d) develop-blue print of a unit test based on a learning unit,
(e) select or construct the questions appropriate to blue-print,
(f) undertake question-wise analysis to check and verify set of questions in terms of blue-print developed,

(g) appreciate the need for different varieties of unit tests to be used as oral or written devices, and
(h) use the different unit tests for various purposes.

1. The Unit Concept

A unit, sometimes regarded as an equivalent to a particular topic of the syllabus, or it represents a combination of two or more topics which are related. However, instead of defining a unit in terms of topics it may be better to understand it in terms of a related learning experience. In that case it may be an organisation of related learning experiences which involve unitary organisation of experiences. Some people try to differentiate between the units of experience and the units of subject matter. It may be difficult to maintain this dichotomy when we regard subject matter as recorded experience and as a source of learning experiences. Nevertheless, one must visualize some source of unity round which learning experiences may be organised. From this point of view, a unit is something more than a mere lesson or period of activity or an arbitrary division of experience. It is in fact an organisation of a number of experiences organised in relation to some of the many possible sources of unity. This source of unity may be in regard to the organisation in planning the curriculum as well as to the classroom development.

Depending upon pupils' natural and social environments, a unit may be titled like, "The town water supply", "The school garden", "The local park", "The city zoo", "A horticulture farm", etc. Likewise, themes, problems, interests, processes etc., may form the basis of unit formation. Selection and planning of the units are in general determined by the design and general framework of the curriculum. The choice of the unit is influenced by the pattern of a curriculum design. In a subject-centred curriculum, the logical unit is the basis for division of the subject involved, whereas it is derived in terms of the maturity of the learners if it is based on the areas of living. If designed on the basis of the experience of the learners, the unit may be organised around cluster points of experiences or problems and interests of learners. Whatever the form of the unit may be, it should be on the integrated area of a chunk of

subject matter which is teachable in approximately a week's time or a fortnight. (1)

2. Concept of a Unit Test

A unit test is a written, oral or practical exercise based on the objectives and content of the unit taught and used as a formal or an informal device for measuring students' achievement in order to improve it through feedback. Thus, a unit test is

(a) written, oral or practical.
(b) formal or informal device of measurement.
(c) based on specific content of a unit.
(d) limited to the testing of specific objectives of the unit.
(e) administered after the completion of the unit.
(f) meant for diagnosis of pupils' achievement through feedback of results.

A unit test is a question paper in miniature. There is not much difference between the two except their scope and purpose. The main purpose of a question paper is that of the measurement of achievement and that of a unit test is improvement of pupils' achievement. A unit test, unlike a question paper, may be mono-objective or multi-objective but a question paper is always multi-objective. The coverage of the content is intensive in a unit test while it is extensive and selective in the case of a question paper. A unit test may or may not include all the forms of questions and may have any number of questions depending upon the length of the unit of teaching. The time is also not fixed. It may vary from 10 minutes to an hour depending on the length of the unit, in contrast to the usual time of two to three hours in a question paper. The results of a question paper are used mainly for grading students. In case of a unit test, the results are used for diagnosis and improvement of achievement through regular feedback.

3. Types of Unit Tests

Unit tests may be of different types. There is no rigid and fast demarcation. Depending upon the purpose of the unit, its scope

and the nature of the subject matter to be tested, unit tests may be prepared in various forms. A unit test may be of 10 minutes' duration or of an hour. However, it is presumed that generally a unit test would not exceed one teaching period. In case of bigger units, more than one unit test may be developed. We may have unit test of different varieties like the following: (2)

(i) A unit test may contain only objective-type and short-answer type questions where the subject-matter is such that it is not possible to cover and test adequately the content through one type of questions and long answer questions are not warranted.

(ii) A unit test may contain only essay-type and short-answer type questions where it may be difficult to have objective-type questions and at the same time is not possible to ignore the essay-type question to test certain abilities like summarisation, integration, expression etc.

(iii) A unit test may contain only short-answer type questions which may range from 5 to 20 in number. Such a unit test may be used when the unit does not demand extended responses with respect to the various learning points and at the same time there is no facility for using objective-type questions. Ease of construction and lack of training of students in attempting objective-type questions could be other considerations.

(iv) A unit test may contain only objective-type questions. Such a test may be usable where cyclostyling facilities are available in the school and the unit to be tested lends itself better to objective scoring. It may also be profitably used where a unit is very long and multiple-choice type questions may serve better for adequate sampling.

(v) A unit test may also be framed sometimes, say every quarterly, to test the various mental processes developed during teaching. In such a unit test, questions of any type may be included but all the specific intended outcomes implied in the objectives are tested. The results of such unit tests are very useful for improving instruction by diagnosing certain abilities which hitherto have not been developed.

(vi) Sometime it is not possible, for want of time, to administer a separate unit test for each unit. In such a case, two or three small units may be clubbed together for testing. This is desirable only when the units are closely related. Apart from the economy of time, it is possible to judge common features and distinguishing features involving related concepts in the units under testing. For example, units of teaching, say, photosynthesis and respiration can be covered by a single unit test.

(vii) A unit test may also be based on a single diagram and questions testing different abilities may be put on that very diagram. Such a test may even be orally used with the class for review purposes.

(viii) Sometimes a unit test may be based on a single table. For example in some unit tests on genetics, a table may be given with serial numbers in each compartment and the students may be asked questions only in terms of serial numbers in each compartment given in that table. Such a unit test is very handy for use and can also be used orally to review a learning unit.

(ix) A unit test may sometimes be prepared in such a way that a simple, two-dimensional chart may be given with some entries in it. Students may be asked to fill in the rest of the entries. Such a unit test may sometimes be used to test basic understanding of biological facts necessary for learning a new lesson. Before taking up new lessons, previous knowledge of the students can be tested by giving such a chart to them for filling up. This chart will be quite useful for understanding the weaknesses of students with respect to the basic facts, which can become basis for remediation before going ahead with the next unit.

(x) It may be necessary sometimes, especially in the case of a unit which by and large is difficult for an average student to follow, that a unit test be framed on selected learning points (concepts) from that unit which the teacher considers very difficult to understand. In such a unit test, the emphasis would be on intensive coverage of those hard spots which the teacher believes the students may have

difficulty to follow. The results of such a unit test may highlight or pinpoint the area of weaknesses and serve as a diagnostic test with limited use, with reference to a particular topic in the class concerned. Such a test would only include questions on the basic understandings required to comprehend the major concepts involved in that unit. True and false statements, cyclostyled on a sheet of paper, can be profitably used in such cases.

4. Blue-Print of a Unit Test

4.1. *Three Dimensional Grid*

Most trained teachers do know what a blue print is, especially those trained in paper-setting workshops conducted by the N.C.E.R.T. or by state educational agencies. Nevertheless, only a few really understand the rationale behind the development of a blue-print and fewer still can apply their knowledge of blue-print to develop a good, workable blue-print. The two-dimensional grid or blue-print, representing the objectives and the sub-units is quite common but the 3-dimensional blue-print, which is an NCERT innovation, is a really useful device to ensure the needed validity and reliability of a unit test. A blue-print as envisaged in the N.C.E.R.T. literature

(a) is a three-dimensional grid,
(b) shows the nature and distribution of questions over different sub-units,
(c) indicates the weightages given to various forms of questions,
(d) pin-points the number of questions with marks, to be set testing different objectives in relation to different sub-units,
(e) shows weightages given to various sub-units,
(f) reflects the emphasis given to various objectives, and
(g) depicts the structure and format of the intended test.

4.2. *The Underlying Assumptions*

While preparing a blue-print, it is assumed that the teacher has already made all out efforts to develop various abilities (as

implied in an objective) and accordingly the weightage to the Application objective, as compared to the Understanding and Knowledge objectives, shows that in teaching more emphasis was laid on the latter two. Another assumption maybe that the unit under testing is less amenable to the Application objective as compared to the other two objectives. Similarly, the assumption underlying differential weightage to various sub-unit is that one sub-unit is more important or has more potential than the others and encompasses more concepts, principles or other content elements than other sub-units. Likewise, the use of one form of questions in preference to another in a particular sub-unit is based on the assumption that the nature and magnitude of the content in that sub-unit demands that particular form to test effectively the desired objective or because of the facility of construction of that variety say, multiple-choice, from that content area. Similarly, the weightage given to different forms of questions assumes that the tester is trying to maintain a reasonable level of scoring objectivity. Overemphasis on a particular variety of question or under emphasis of another variety in relation to another sub-unit is based on the assumption that the content elements of that unit are more amenable or less manipulative to that variety to test the desired objective as indicated in the blue-print.

Thus a blue-print, when prepared, does involve certain assumptions based on the limitations and uncertainties of the testing possibilities. Some of these assumptions can be questioned by another teacher evaluator and, therefore, he can develop another blue-print on the same unit, which may be quite different in format and structure from the proposed unit test. Before starting a blue-print, it is thus essential that the teacher becomes cognizant of the content elements of that unit to judge the nature and scope of the unit. This ensures better distribution of questions of different types to test various objectives in relation to different sub-units delineated at the beginning.

4.3. *Methodology of Preparing a Blue-Print*

Looking at any of the ready-made blue-prints may give as the idea that it is only filling-up of the boxes according to one's

(tester's) will. In fact it is not so. Any blue-print which is prepared without cognising the intended anatomy of a unit test is sure to create difficulties while framing the questions. An application question inserted against a sub-unit might be difficult to frame because the content elements therein do not lend themselves such a question. Likewise, a multiple-choice question shown against a particular sub-unit testing a specific objective may not be possible but a short-answer question may be easier to frame for the same purpose. Thus while making entries during blue-printing, care must be taken that one has a clear picture of the problems and possibilities of framing a particular form of question in a sub-unit testing the intended instructional objective.

For getting an insight into the architecture of a blue-print, a lot of practice is needed. Though no specific rules can be given for the methodology of developing a blue-print yet on the basis of my experience. I venture to provide the following steps as guidelines for developing a blue-print, assuming that the teacher or evaluator is familiar with the content elements of the unit concerned and also the instructional objectives which are testable in a written test i.e. Knowledge, Understanding, Application and Skill (only drawing) and the unit can be taught within a week or two. For illustration purpose, a sample blue-print on a unit, "a cell", for class IX based on a State syllabus is given in the form of an annexure for reference.

Steps:

(a) Delineate the unit into sub-units keeping in view an integrated chunk of content which can be taught almost independently as a module. Depending upon the nature and size of the unit, we can divide a unit into three to five or six sub-units. Less than three would mean one may not be able to ensure adequate coverage of various content elements as it quite likely that the tester may overlook significant content elements which is not possible when 4, 5 or 6 sub-units are identified. On the other hand, if we have more than 5 or 6 sub-units it becomes quite cumbersome, like a full-fledged question paper. Moreover, this leads to low weightages to various sub-units which might prevent

inclusion of essay-type questions (when content demands it) which usually carry not less than 4 or 5 marks. Thus from the point of view of practicability and to ensure the needed coverage of content, 3 to 5 sub-units may be considered desirable. (5 in the sample)

(b) Give proportionate weightage to various sub-units depending upon the significant facts, concepts and textual details forming the magnitude or bulk of content in that sub-unit. (2, 5, 7, 6, 5 in the sample)

(c) Formulate or identify the objectives (K-U-AP-SK) which can be tested using the content of the unit. All the four objectives may not be necessary for every unit. Some may demand two, others three and some only one. (K-U-AP and SK in the sample)

(d) Assign proportionate weightages to each of the objectives identified under step-3. (8, 7, 5, 5 in the sample)
Weightage to various objectives depend on:

(i) amenability of content to testing of different objectives.
(ii) the use to which the test results will be put. If a remedial programme is to be organised then more weightage to the knowledge objective has to be given or it may be an exclusively knowledge-based unit test.
(iii) previous performance of students on these objectives.
(iv) intended learning outcomes to be used as criteria or standard of performance.

(e) Decide the approximate time to be alloted for the unit test. It can vary from 10 mts. to 1 hour depending upon the size of the unit. However, one teaching period may usually be considered appropriate for a unit involving 5 to 10 teaching periods. (40 mts. in the sample)

(f) Decide the total marks to be alloted to the unit test. This would depend on the number and form of questions to be included. However, in reality we have to fix-up first the time limit arbitrarily, say 30-40 mts. and then decide the number of questions (25 in the sample).

(g) Decide the number and form of questions to be included in the test (E, S.A., V.S.A., M.C. in the sample) and give

proportionate weightage, depending upon the nature of the content and familiarity of students with one or the other form. A unit test may have one or more type of questions with some variation in the weightages given to each type. (5, 10, 5, 5 in the sample)

(h) Now start filling in the boxes of the grid as under: (Refer to the sample in the annexure):

(i) Insert the total marks in the bottom right-hand corner box as indicated by (a) in the blue-print. Enter sub-unit-wise marks and objective-wise in columns (a^2) (a^1) in rows.

(ii) Then decide about the essay-type questions which could be given only where weightage to a sub-unit is sufficient enough and the content demands a long-answer question. Moreover certain essay-type questions do include a part requiring students to draw a diagram as a part of the same question. In such cases, set the marks apart for the skill part of the essay question under the skill column in the same row against the same sub-unit as shown in sub-unit-3 (3 marks for skill). This helps the teacher to apportion the remaining weightage of the skill objective to other question/s to be set, testing skill. The figures inside the bracket indicates the number of questions to be set and the figures outside indicates the total marks for the question/s. In the sample unit against sub-unit-3, under essay-type, 2 marks+3 marks for K and SK are shown. Since the skill part is included in the same question under essay type, the number of question/s is not shown e.g. 3(−), in column 'E' under skill objective, so that only one question is represented (indicated by (b) in the blue-print).

(iii) The next step is to fill-in the skill column so that whichever sub-unit is more amenable to testing of the drawing skill, could be used for alloting skill-based questions. Sometimes it is necessary to frame a full-fledged question testing skill independently. This is possible only if we first explore the possibility of identifying the sub-unit which admits the

skill-based question as indicated against sub-unit-2 under S.A. the column of skill-objective. (indicated by 'c')

(iv) Once skill questions are inserted, procede to allocate Application questions. The simple reason for this is the inherent difficulty of framing such questions on every sub-unit. Therefore, we should identify the relevant sub-unit, having more potentiality than others for framing questions testing the application objective. This prevents the difficulty that may arise later in converting Knowledge or Understanding question to the Application type question which is not always possible. Likewise content elements in that particular sub-unit/s may be more amenable to multiple-choice than the short answer variety of question for testing the Application objective. Sometimes all the sub-units could be used for testing the Application objective although they may not admit all forms of question. (indicated by 'd' against sub-unit 2 under Application Objective).

(v) Once the distribution of Skill and Application questions is over, the task becomes easier as conversion of Knowledge-based questions to the Understanding type or vice-versa is no problem. Nevertheless, it is suggested that we may now procede sub-unit-wise making entries simultaneously under Knowledge and Understanding objectives. Care has to be taken that the decision about allocating objective-type questions should precede short-answer questions. Likewise distribution of questions under the objective 'Understanding' should precede the Knowledge objective so that choice of content element could be made for higher level objectives in precedence to lower level objectives. Thus, entries in all the sub-units are made sequentially as shown in the sample test. (marked as e, f, g, h, i and j). Thus (a) to (j) indicates sequential steps for developing a blue-print.

Distribution of questions over K and U against each

sub-unit is given below:

Sub-unit-1	U—V.S.A.	1(1)	K—V.S.A.	1(1)
Sub-unit-2	U—M.C.	1(1)	K—	Nil
Sub-unit-3	U—	Nil	K—V.S.A.	1(1)
Sub-unit-4	U—S.A.	2(1)	K—S.A.	2(1)
	M.C.	1(1)		
Sub-unit-5	U—S.A.	2(1)	K—V.S.A.	1(1)
			M.C.	1(1)

(vi) More often than not, some adjustment about the distribution of questions under the Knowledge and Understanding objective, becomes necessary in one sub-unit or two to ensure the correct weightage to sub-units and objectives. This is inevitable unless one deliberately goes ahead keeping in view only the adjustment of marks and not the content elements of the sub-units.

(vii) Insert row-wise and column-wise sub-totals indicating the number of questions and the marks for those questions. Check the totals on both the sides to see that it is the same (25 in the Sample). The blue-print is now ready.

(viii) Check whether in the blue-print, the questions are well spread over the various sub-units and are not crowded at one place. See that by and large different units are not restricted to one type of questions only although in a particular sub-unit it might be inevitable.

4.4. *Using a Blue-Print*

Once the blue-print is ready, framing of questions on each sub-unit becomes possible. It provides the basis for framing of different types of questions, testing different abilities, using a specific content. Thus the construction of each item is focussed on a pre-determined objective, thereby ensuring that all questions are objective-based and cover the content adequately. More than one blue-print can also be prepared on the same

unit. Two unit tests prepared on the same unit of teaching when administered to the same class can give an idea about the validity of students' performance. To what extent performance of students on the two units go together would indicate the degree of equivalence of the two tests. If two or three questions are framed on each of the questions indicated in the same blue-print, one can develop parallel unit tests. If certain absentees or remedial-group students are to be retested on that particular unit, the parallel unit test prepared on the same blue print can be profitably used. The results may be used for determining the reliability of the unit test when two versions of the same unit test (based on the same blue print) i.e. parallel unit tests, are administered at the same time.

A study of different blue-prints prepared by a teacher from time to time on various units of the syllabus would throw light on the instructional strategies of the teacher regarding emphasis be laid on the various objectives and the type of questions he used. Do his/her blue-prints reflect almost the same weightage to instructional objectives on all units?Do higher level objectives receive more attention towards the later part of the session? Does he/she use only a particular form of question or practises all forms of questions? Thus the study of blue-prints used by a teacher over a period of time gives an insight into the emphasis laid by the teacher on various instructional objectives. If blue-prints prepared by different teachers for a particular unit test or on various units are compared, interesting inferences can be drawn on the basis of which guidelines can be provided to the teachers for developing blue prints envisaging comparability of students' performance.

The blue-print of a unit test provides a scientific basis for the construction of a unit test. It takes care of adequate coverage of content and objectives to ensure the needed validity. Questions developed on such blue-prints by different teachers or by the same teacher over a long period can ultimately lead to the development of a good question bank comprising objective-based questions on different units of the syllabus. When this stage comes after a good deal of practice of framing questions on pre-designed blue prints, the necessity of developing blue prints for every unit may not be felt. The ultimate aim of

developing the blue print is to take cognizance of the validity and reliability of the unit test and not merely filling in of the boxes with any type of questions under any objective and sub-unit.

5. Framing of Questions

Once the blue print of the Unit Test is ready the next step is to construct questions of various types, testing different objectives as indicated in the blue print against sub-units or topics. The quality of these question is to be maintained as discussed in the previous chapter.

6. Editing and Consolidation

Each question needs to be edited before it is finalised. This editing includes ensuring relevance of the intended answer with the model answer prepared at the time of construction of the question. This also includes verifying the key (answers) as well as the format of the question besides checking the relevance of each question testing the intended objectives. This is followed by consolidation of questions in a pre-determined order which may be in terms of content, objectives, form of questions or difficulty level of questions.

7. Instructions for the Examinees

When the unit test is ready general instructions are developed which provide the needed information and guidelines for the examinees to take the test wisely.

A sample unit test along with a blue print is given in the Annexure-A.

8. Using the Unit Tests

Since evaluation is an integral part of teaching and learning, students are observed in various situations continuously where measurement of their achievement may be possible. For this purpose, various tools and techniques are u sed. The unit test is one of such tools.

In a yearly plan of instruction, a teacher has to divide the prescribed syllabus into various learning units. A learning unit comprises such topics which have some sort of continuity and coherence. While deciding about a teaching unit, the teacher has to take into consideration the achievement of the class, the time factor and the emphasis given to the topics and the weightage given to it in the syllabus. He has to determine the instructional objectives of the unit also. When the instructional programme of this particular unit is over, the teacher has to appraise the students' achievements in terms of the instructional objectives specified at the beginning. For this purpose he has to develop a formal or informal test which may be called a unit test for that particular unit. Such unit tests when properly constructed and used, serve a number of useful purposes, some of which are given below: (4)

(i) *For measuring achievement*

Such tests can be used as classroom tests for testing the achievement of students. These may be administered after completion of every unit of teaching. This enables the teacher to know the rate of progress of students as well as the extent of attainment of the objective of the unit.

(ii) *For diagnosis of students' weaknesses*

The analysis of the test result's may provide evidence of poor achievement in certain areas by an individual or a group of students. The teacher can identify weak students and areas of their weaknesses. This helps him to adopt remedial measures for the individual or the group to facilitate further learning. If need be, he can frame a diagnostic test on the hard spots of learning to find out the exact nature of difficulties. (5)

(iii) *For accelerating achievement*

On the basis of the test results, teaching can also locate certain areas of achievement where better achievement by the students is possible. The teacher can then concentrate his efforts on an individual or a group where achievement of excellence in that area would be possible. In fact, exceptionally high achievement of certain students gives a clue to the need for an enrichment

programme.

(iv) *For improving instruction and efficiency*
The tests not only help a teacher in locating the strong and weak points of the students but also help in discovering the strong and weak points of his methodology, his understanding of a certain concept. A defective method adopted by teacher becomes recognizable by the negative response from the students. In such cases, the teacher can modify his method of teaching or strategy of unit planning. There can be other situations where even backward students are able to understand and grasp some difficult points because of the use of more appropriate methods. Here the teacher can develop confidence in his methods.

(v) *For motivating the learners*
If unit tests are given at regular intervals, they act as good incentives for students to evaluate their progress regularly and also to improve their performance further. They also motivate students to develop regular study habits. A well-planned use of unt tests at regular intervals discourages selective study on the part of students and provides hints for remedial measures.

(vi) *For instructional purpose as teaching aids*
Various questions of the unit-tests can be used as good teaching devices. Here short-answer type questions are of special significance for use on various occasions, before, during or after instruction. Essay and objective type questions, especially the multiple-choice type questions can also be used profitably if these questions are thought-provoking. Especially, short-answer type questions lend themselves better for use as teaching devices. These can be used in pre-testing, development of a lesson, review of a lesson and for diagnosing, in a rough and ready manner, the difficulties of pupils. (6)

(vii) *For pupil's assignments*
Instead of using a unit test as an instrument of evaluation, it can be used in full or in part as home assignments. It is convenient for the teacher to pick up some questions from the unit tests available and use it as a school or home assignment.

As unit test questions are well chosen, in accordance with the requirements of the unit, the teacher need not frame fresh questions everytime. He can make use of selected questions. Some of the questions can be even utilised for undertaking small projects.

(viii) *For students' self-evaluation*
As every unit test carries with it a key for objective-type questions and outline answers, for each short answer and essay-type question it can be used by the student for self-evaluation. But this is possible only when the school maintains a library of unit tests and a question bank which is available to the students. When available, students can pick up questions from the unit test, write answers and compare them with the given key or outline answers. This helps them to know their strengths and weaknesses at any time.

(ix) *For developing a question bank*
When a unit test on each unit of the syllabus is prepared by the teacher, he will have, in course of time, a good number of unit tests on all the units of teaching. This will lead to the development of good pool of unit tests. This pool will continue to be argumented year by year and up dated from time to time. Each question from these unit tests can be transferred. Such a pool can be taken advantage of by all subject teachers for use in their evaluation programme on specially prepared cards of about 5″×8″ size. Such cards have question on one side and key or outline answer on the reverse side. By and by a good pool of questions on these cards is developed. Such pool of unit tests and individual questions can be developed for different subjects for different classes. There can be more than one unit test developed on the same unit. Keeping in view the limitation of time and needed technical competence for construction of good questions, need for supply of ready made questions to the teachers in the form of item bank or unit tests library cannot be over-emphasised.

(x) *For reinforcement and feedback*
Continuous use of unit tests by teachers and reporting of results

to students is essential. Results of unit tests should be communicated as early as possible. This serves two purposes, one of reinforcement of learning, and the other of feedback for teachers and students both. Continual feedback of unit tests helps in acquainting students about the adequacies and inadequacies in their learning. This provides motivation on the one hand and develops positive self-concept on the other by recognising their achievement.

ANNEXURE

Subject : Biology Unit : Cell Class : IX Time : 40 Mts. Max. Marks : 25

Blue-print

Objective	Knowledge				Understanding				Application				Skill		Total
Form of questions →	E	SA	VSA	O	E	SA	VSA	O	E	SA	VSA	O	E	SA	Total
S. No. Content Units ↓															
1. Cell Theory	—	—	1 (1) (e)	—	—	—	1 (1) (e)	—	—	—	—	—	—	—	2 (2) (a^2)
2. Structure of Cells	—	—	—	—	—	—	—	1 (1) (f)	—	2 (1) (d)	—	—	—	2 (1) (c)	5 (3) (a^2)
3. Cell Organelles under E. Microscope	*2 (1)	—	1 (1) (g)	—	—	—	—	—	—	—	—	1 (1) (d)	*3 (—) (b)	—	7 (3) (a^2)
4. Cell Division Mitosis	—	2 (1) (h)	—	—	—	2 (1) (h)	—	1 (1) (h)	—	—	—	1 (1) (d)	—	—	6 (4) (a^2)
5. Cell Division Meiosis	—	—	1 (1) (i)	1 (1) (i)	—	2 (1) (i)	—	—	—	—	1 (1) (d)	—	—	—	5 (4) (a^2)

Sub-Total	*2 (1)	2 (1)	3 (3)	1 (1)	—	4 (2)	1 (1)	2 (2)	—	2 (1)	1 (1)	2 (2)	*3 (—)	2 (1)	25 (16)
Total		8					7				5		5		25
		(a^1)					(a^1)				(a^1)		(a^1)		(a^1)

Notes: Figures within brackets indicate the number of questions and figures outside the brackets indicate marks.

*Denotes that marks of two parts of the same question have been combined.

Summary

Essay (E)	No. 1	Marks: 5
Short Answer (SA)	No. 5	Marks: 10
Very Short Answer (VSA)	No. 5	Marks: 5
Objective (O)	No. 5	Marks: 5
	Total = 16 questions	Total = 25 marks

Sample Unit Test

Unit : **Man and Food Supply** **Class : X**

Maximum Marks : 30 **Time : 40 mts.**

BLUEPRINT

Objective	Knowledge				Understanding				Application				Skill		Total
Form of Questions / Content Units	E	SA	VSA	O	E	SA	VSA	O	E	SA	VSA	O	E	SA	
1. Population distribution	—	—	—	1(1)	—	2(1)	—	—	—	2(1)	—	2(2)	—	—	7(5)
2. Population growth	5(1)	—	—	—	—	2(1)	—	—	—	2(1)	—	—	—	—	9(3)
3. Food webs	—	—	—	1(1)	—	2(1)	—	2(2)	—	2(1)	—	1(1)	—	—	8(6)
4. Food supply	—	—	—	(1)	—	2(1)	—	—	—	2(1)	—	1(1)	—	—	6(4)
Sub-Totals	5(1)	—	—	3(3)	—	8(4)	—	2(2)	—	8(4)	—	4(4)	—	—	30(18)
Total	8				10				12						30

Notes : *Figures within brackets indicate the number of questions and figures outside the brackets indicate marks.*
**Denotes that marks have been combined in the same question testing two different objectives*

Summary	No. of questions	Marks allotted
Essay [E]	1	5
Short Answer [SA]	8	16
Very Short Answer [VSA]	0	9
Objective [O]	9	—
Total	18	—
		Total 30

Scheme of Options : No option is provided.
Scheme of Sections : Only one section.

MAN AND FOOD SUPPLY

Class : X

Time : 40 mts.
M. Marks : 30

Instructions

I. All questions are compulsory.

II. Answers are to be written in the answer book.

III. Questions 1 to 9 have 4 suggested answers A, B, C, and D only one of which is correct. The letter indicating the correct answer should be written in the answer book. Each carries one mark.

IV. Questions 10 to 17 carry 2 marks each and answers should not exceed 10-15 words.

TEST

1. Which of the following trends of population growth rate is seen in Western Europe?
   A. Births equal deaths
   B. Births lower than deaths
   C. Births higher than deaths
   D. None of the above

2. Which of the following is the major function of a food chain?
   A. Food supply to other organisms
   B. Constant supply of energy
   C. Change in types of foods
   D. Energy transfer from lower to higher organisms

3. Urban population in developing countries increases because of
   A. higher birth rate in cities.
   B. rural to urban migration.
   C. greater opportunity of urban employment.
   D. lower death rate in cities.

4. The relationship between green plants and producers is similar to the relation between bacteria and

A. producers.
B. consumers.
C. decomposers.
D. nutrients.

5. Productivity of forest land is higher than cultivated land because

A. cultivated lands have loose soil.
B. forests replenish nutrients.
C. cultivated lands are tilled.
D. forests retain moisture.

6. To check the decrease in the density of population where mostly *adivasis* are concentrated which of the following important measures would you take?

A. Bring more land under cultivation by clearing forests.
B. Build huts and roads for *adivasis*.
C. Open dispensaries and schools.
D. Increase land under settled agriculture.

7. *Statements*

I. As technology improves man's ability to adapt his environment increases.
II. The percentage share of employment in resource industries like grazing, mining, farming are shrinking in the U.S.A.

Direction
Indicate which of the following BEST describes the relationship between the two statements given above?

A. The situation in statement II contradicts the principle in statement I.
B. The situation in statement II neither contradicts nor can be explained by the principle in statement I.
C. The situation in statement II can be explained by the principle in statement I.
D. The situation in statement II is consistent with statement I.

8. Productivity is lower in the grasslands than in the forests because of

A. less producers than consumers.
B. less producers than decomposers.
C. more producers than consumers.
D. more producers than decomposers.

9. Which of the given changes in food habits will favourably affect capacity of land supporting a number of people?

A. Less consumption of plant and animal foods.
B Equal consumption of plant and animal foods.
C. More consumption of plant foods.
D. More consumption of animal products.

10. 'Future pattern of high density of population areas will be largely dispersed all over the world'. Why?

11. Population growth rate has been 0.1% per annum with the practice of settled agriculture. Why had this growth rate increased only after the onset of settled agriculture and not prior to it?

12. Why is grazing on open pastures less productive for the rearing of cattle for beef? What is the better method which is practised now?

13. Explain why commercial fishing has become the main economic pursuit of N.W. Europe, Japan and China over beef cattle rearing.

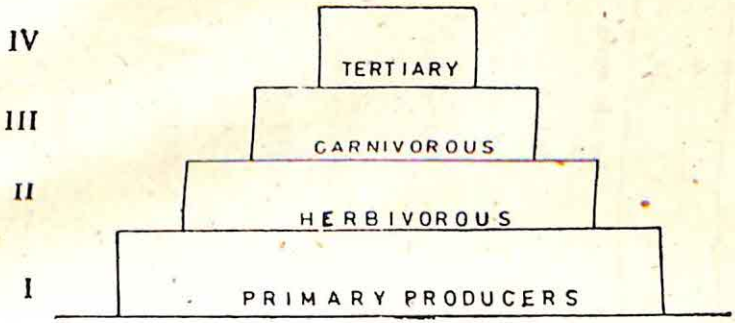
14. Distribution of population in terms of number per unit land is not the only aspect of population studies. Suggest two other aspects which may also explain the real phenomenon of population distribution.

15. **The world is finite, resources scarce, oil is going, ores depleted. Man is far too enterprising. Fire will rage with man to fan it. Soon we'll have a plundered planet. People breed like fertile rabbits, people have disgusting habits.**
Moral
The evolutionary plan
went astray when evolving man.

—Kenneth Boulding

Suggest two measures to this conservationist's complaint to preserve our planet from being depleted.

16.



In the pyramid of numbers given above which level will have to be reduced in order to ensure an ecological balance of producers?

17. Read the statements given below and answer the question.

I. Rate of increase of 3 per cent is necessary to maintain an adequate level for the entire world population in 2000 AD.

II. Per capita income improvement is minimal in the last decade, therefore, deficiency gap continues to widen as one enters the stage of maximum population growth.

What is the main difference in the approaches of these two hypotheses which explain the effect of population growth on economic growth?

18. What are the main trends of population growth rates in 1. Australia, 2. Asia, 3. Europe? How does this affect the age structure of population in the economically productive age group in these countries?

Scoring key and Marking Scheme

Q. No.	1	2	3	4	5	6	7	8	9
Key	D	B	C	C	B	D	A	A	D

Q. No.	Expected outline answers	Part-wise marks
10.	(a) With the opening up of new areas. (b) Availability of new resources and occupations.	2×1=2
11.	(a) Lowering of death rate. (b) More births consequently and increase in population growth.	2×1=2
12.	(a) The beef cattle on open pastures do not increase the productivity of soils as there is no return of nutrients. (b) It is better to grow the crops to feed the animals.	2×1=2
13.	(a) Lack of arable land and low productivity of land (b) Availability of long coastal fishing areas.	2×1=2
14.	(a) Information on urban and rural population break-up in the same regions. (b) Male-female break-up	2×1=2
15.	(a) Use of potential resources (b) Efficient technology to reduce wastage (c) Improvement of economic gains to man—(Any two of these)	2×1=2
16.	(a) Carnivorous (b) Tertiary consumers	2×1=2

17. *Statement A*
Denies the possible role of meeting any more population growth.

Statement B
Analyses the impact of economic growth on supplying sufficient food for people.

$2 \times 1 = 2$

		Growth rate	*Work force*
18.	(a) Australia	: Low birth rate Low death rate	—Few persons who are young and economically productive
	(b) Asia	: High birth rate Low death rate	—Young population with few workers in the productivity age group
	(c) Europe	: Low birth rate High death rate	—More persons are young and in the productive working age group

(a) $2 \times \frac{1}{2} + 1 =$
(b) $2 \times \frac{1}{2} + 1 =$ 5
(c) $2 \times \frac{1}{2} + 1 =$

any 10 correct points out of 12

(i)
(ii)
(iii)
(iv)
(v)
(vi)
(vii)
(viii)
(ix)
(x)
(xi)
(xii)

Question-wise Analysis

S. No.	Objective	Specification	Content Unit	Type of questions*	Marks allotted	Estimated time for answering	Estimated difficulty level**
1.	Knowledge	Recalls	Population distribution	O	1	1	C
2.	Knowledge	Recalls	Food Webs	O	1	1	C
3.	Knowledge	Recalls	Food Supply	O	1	1	C
4.	Understanding	Compares	Food Webs	O	1	1	B
5.	Understanding	Identifies	Food Webs	O	1	1	B
6.	Application	Infers	Population distribution	O	1	1	A
7.	Application	Infers	Population distribution	O	1	1	A
8.	Application	Establishes relationship	Food Webs	O	1	1	A
9.	Application	Establishes relationship	Food Supply	O	1	1	A
10.	Understanding	Identifies	Food Supply	SA	2	2	B
11.	Understanding	Compares	Food Supply	SA	2	2	B
12.	Understanding	Identifies	Food Supply	SA	2	2	B
13.	Understanding	Identifies	Food Supply	SA	2	2	B
14.	Application	Suggests	Food Supply	SA	2	2	A

15.	Application	Suggests	Food Supply	SA	2	2	A
16.	Application	Establishes relationship	Food Supply	SA	2	2	A
17.	Application	Judges	Food Supply	SA	2	2	A
18.	Knowledge	Recalls	Food Supply	E	5	15	C
					Total 30 marks	40 mts.	

*O Objective
VSA Very Short Answer
SA Short Answer
E Essay

**A for Difficulties
B for Average
C for Easy

CHAPTER XIV

TECHNOLOGY OF SETTING BETTER QUESTION PAPERS

1. Introduction

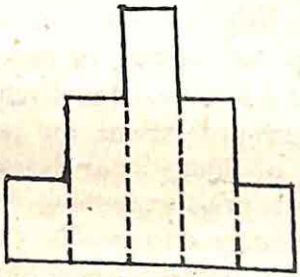
One of the most commonly used techniques for measuring cognitive learning is the written examination. Written tests are, therefore, the main instruments of evaluation used in public examinations as well as classroom testing. This is true for almost all the countries of the world. Written tests may take the form of periodical tests, term tests and annual tests. Each one of these tests in heavily loaded with verbal content and can be scored quantitatively on a 0 to 100 points scale. Compared to oral and practical tests, written tests are much easier to construct, administer and score. Moreover, the tester can ensure a good deal of objectivity in scoring which makes them a more valid and reliable measure of students' learning. This is precisely the main reason why written examinations have come to stay and have attracted more attention of the educators and researchers than have oral or practical examinations. In this chapter attempt is made to enable the reader to

(a) recall the various steps in setting better question papers.
(b) design the question papers to ensure maximum validity.
(c) develop blue print of a question paper on the basis of a given design.
(d) select or construct the appropriate forms of questions in conformity with the blue print.
(e) develop scoring key and marking schemes for each question of a question paper for improving scoring objectivity.
(f) edit the question paper through question-wise analysis.
(g) judge the given question paper and suggest improvement.

2. Need for Cognizance of Objectives

Inspite of the fact that written tests have been in vogue for centuries, it was only during the fifties of this century that the need for objective-based testing was realised. The main impetus to recognise instructional objectives as the basis for teaching and testing was given by Bloom's taxonomy (1) of Educational objectives: Cognitive domain (1956). Thereafter a lot of work on objectives has been done. In India too, this philosophy of behaviourism caught the fancy of the researchers and curriculum workers. In fact, a lot of work on objective-based testing has been done by the National Council of Educational Research and Training, New Delhi, since the early sixties.

In order to improve the validity of written tests, the need for formulating instructional objectivities in advance cannot be overemphasised. Objectives in the cognitive domain are generally stated in terms of Knowledge, Comprehension, Application, Analysis, Synthesis and Evaluation according to Bloom's taxonomy. However, in India the NCERT has developed its own taxonomy of instructional objectives after working with thousands of teachers and paper setters who were involved in evaluation courses organised for the training of paper-setters. Keeping in view the researches on Bloom's taxonomy and the difficulty of an average teacher in comprehending and using this taxonomy, the NCERT has worked out a more realistic and workable taxonomy (2) which is now in use in almost all the states of India. This is a three-tier taxonomy comprising Knowledge, Understanding and Application objectives arranged in order of hierarchy. Each of these objectives is further specified in terms of expected learning outcomes. The three objectives which form the basis of testing cognitive learning are depicted in the following sketch which has become almost a cliche in



the field of written examinations. From the diagram it is

evident that three major objectives of biology teaching have been cognised, viz. knowledge of biological facts and phenomena, understanding of biological concepts and principles and their application to new or unfamiliar situations. All the three are depicted in order of complexity. The knowledge objective is defined into two, understanding into eight and application into seven specifications or expected learning outcomes. It is assumed that the knowledge objective is subsumed under objective understanding which in turn is subsumed under the application objective. Thus a question on understanding automatically tests knowledge as well, while an application question tests understanding *vis-a-vis* knowledge also. All these objectives of biology are stated in behavioural terms as given in the Appendix-A.

The validity of a test is further maintained by basing questions on these specifications which are adequately covered during testing. Moreover, proportional weightage is given to the three objectives depending upon the previous learning of students and considering the difficulty level at which a test is to be developed. Thus the role of objectives is very significant in the assessment of cognitive objectives of learning in biology.

3. Designing a Question Paper

Though planning a question paper is considered a routine activity for teachers, yet analysis of teacher-made tests reveals that the setting of question papers is far from satisfactory especially when considered in terms of coverage of instructional objectives, sampling of content, wording of questions and scoring objectivity. This is true of both external as well as internal examinations in vogue. Nevertheless the need for well-designed question papers is now increasingly felt. In fact all the Boards of Secondary Education/School Educations in India have become cognizant of the need for training of paper setters for making external examinations a more valid and reliable measure of student's achievement. Such programmes are regularly organised by the National Council of Educational Research and Training in collaboration with the various examining boards in India. These training courses are conducted to develop expertise in developing good question papers and unit tests. The emphasis in such courses is on objective based testing. The usual pattern of designing question

papers *vis-a-vis* unit tests for written examinations is reflected in Boards, question papers (3).

Basic Requirements

The basic requirement of any given instrument of evaluation is to ensure its validity, reliability, objectivity and practicability to the maximum extent. Improvements relate to the administrative as well as academic features of paper setting. As far as administrative aspects are concerned, the decision is to be taken about the total time for the question paper, the total marks, the number of questions, the scheme of sections, scheme of options etc. As far as academic steps are concerned, these relate to the coverage of objectives, coverage of content, form of questions, quality of questions, the arrangement of questions, marking scheme, model answers etc. Each of these aspects is explained here-in-after.

(i) *Total time*

At the secondary or senior secondary stage, the time allotted for a question paper varies from about 2 hours to 3 hours while in the case of a unit test, it may vary from 15 minutes to 45 minutes depending upon the length of the unit and the teaching period. But a decision about the total time is the first requisite for designing a question paper.

(ii) *Total time*

Although on the face of it total marks do not seem to have any relevance to the quality of the question paper, yet it becomes necessary to allot sometimes more marks or less marks, depending upon the number of questions to be used. The more the marks reserved for a question paper, the easier it becomes to develop detailed marking schemes, thereby ensuring more objectivity in scoring the scripts. With less marks, the possibility of including eassy-type questions decreases and the need for having more short-answer and objective-type question increases.

(iii) *Sections*

Whether a question paper should have one or more sections

also depends on the type of questions to be included. If multiple-choice questions are introduced, a separate time limit may be necessary. But if it is a unit test with very few multiple-choice questions, there is no use having a separate time-limit for these questions. Moreover, from the administrative point of view, it becomes rather difficult to ensure effective separate administration of two types of tests. Sections are sometimes made content-wise, in which case it is advantageous to put a restriction on attempting a fixed number of questions from each section in order to discourage selective study. Likewise, from the question can also be the basis for various sections and it does facilitate scoring besides analysis of results for feedback.

(iv) *Scheme of options*

In most examinations free over-all option is discouraged because of the non-comparability of results. However, internal options of the either-or type are acceptable with a proviso that optional questions should be balanced in terms of the objective tested, the content area, the length of expected response and the difficulty level of the question.

The following aspects of design relate more to the academic side of the question paper.

(v) *Coverage of Instructional Objectives*

Knowledge, understanding, application and skill (drawing skill only) objectives are usually tested in the written examinations. These objectives are given proportionate weightage for the purpose of paper setting. The weightaghes may vary from one examing board to the other depending upon the variation, the previous background of students on these objectives, emphasis on the intended learning outcomes and the difficulty level of the question paper as a whole. This weightage, when applied to unit testing, would vary from unit to unit depending upon the nature, scope and content of the unit. One of the designs of the question papers reveals the following weightages given to different objectives in annual examinations.

Weightages to Objectives

Weightage	OBJECTIVES				
	Knowledge Objective	Understanding Objective	Application Objective	Skill Objective	Total
Marks	20	15	10	5	50
Percentage	40	30	20	10	100

(vi) *Coverage of content/syllabus*

To ensure adequate coverage of content and better sampling of questions on each area, the whole syllabus is divided into convenient units numbering about 8 to 12. Each of these units is given proportionate weightage depending upon the nature, scope and significance of the content-elements. This weightage is determined in terms of its length, place in the textbook, importance of the topic etc. The number of units should neither be too small nor too many. A small number may lead to the neglect of certain topics while too many units may hinder the inclusion of long-answer questions that carry more marks. This weightage is represented in a table like the following:

Weightage to Content Units

Weightage	UNITS OF THE SYLLABUS										
	Unit 1	Unit 2	Unit 3	Unit 4	Uuit 5	Unit 6	Unit 7	Unit 8	Unit 9	Unit 10	Total
Marks	2	4	6	9	5	8	4	3	8	1	50
Percentage	4	8	12	18	10	16	8	6	16	2	100

(vii) *Decision about the Form of Questions*

The form of questions normally depends on the acquaintance of students with the form of questions, the extent of reliability acceptable, the needed objectivity in scoring, the nature of the content and total time stipulated for a test.

Essay-type questions are more frequently used than short answer and objective type questions in most examinations. However, in India, there is now a clear trend of increasing the number of short-answer and objective-type questions with a

corresponding reduction in essay-type questions. The weightage to different forms of questions as revealed in one of the question papers is as follows:

Weightage to Form of Questions

Weightage	Types of questions			
	Essay type	Short answer type	Objective type	Total
Marks	20	20	10	50
Percentage	40	40	20	100

(viii) *Difficulty Level*

In designing the question paper, the expected difficulty-level is also to be stated. Although the actual difficulty-level can be determined only after administration of the test, yet based on previous experience and knowledge experienced teachers and paper setters can well estimate the difficulty level of various questions *viz-a-viz* the whole question paper. In the external examinations where there is a heterogeneous group consisting of bright, average and poor students, the difficulty-level that could normally be distributed is shown in the following table:

Weightage of Difficulty-Level of Questions

Weightage	Estimated Level			
	Difficult	Average	Easy	Total
Marks	10	30	10	50
Percentage	20	60	20	100

For classroom testing, the difficulty-level would depend on the objectives to be tested, the previous standard of the class, the nature of content and the homogeneity or heterogeneity of the group.

All the tables given above are only suggestive and hypothetical and are to be prepared keeping in view the requirements of a particular question paper.

All the above-mentioned features become a part of the design of the question paper which indeed reflects the policy

about paper setting. For external examination, this has to be quite rigid, although broad-based, and should normally continue for three years or so as long as there is no change in the syllabus. The paper-setter has no right to make any change in the design of the question paper but, based on this design, he is free to develop his own blue-print about which a discussion follows.

4. Preparation of the Blue Print

Having received the design of the examining agency, a paper-setter is expected to prepare a blue-print of the intended question paper. This blue-print is a three-dimentional chart, indicating the distribution of various forms of questions in relation to various unit of the syllabus, testing different objectives. This blue-print also depicts the numerical weightage to each question as well as the scheme of options, if any, to be used in the setting of the question paper. This blue-print is the monopoly of the paper-setter. However, this blue-print must be developed within the constraints of the given designs which reflects, in fact, the policy statement of the question paper. Based on one design, many blue-prints can be prepared and the same blue-print can be used for developing more that one question paper which may be used for supplementary examination. Technique of preparing a blue print is discussed in details in Chapter XIII on unit testing.

5. Framing of Questions Based on Blue-Print

Framing of questions may be taken upon each sub-unit or one particular form of question, say, the objective-type may be taken up first, followed by short-answer and essay-type questions. Each question located within a blue-print must reflect the same objective, content and the form of the question. Of course, mastery of the subject matter knowledge of objectives and their specification and the skill in framing different forms of questions are essential. Whatever the type of questions framed, it must be based on a well-defined, pre-determined specific objective relating to a specific content-element and has to be set

at the designed estimated level of difficulty. Each question must satisfy the conditions of its structure. Thus, every question is to be framed only in terms of its position indicated in the blue print.

In hardly needs mention that model answers, alongwith the marking-scheme must be developed at the same time along with the framing of a question.

6. Editing and Consolidation

Once the questions are ready, the next step is the consolidation of those questions. Questions may be assembled into various sections which may be done on the basis of their format. Free-response questions may be at one place while the objective-type at another place. Within each format of question, the same order may be followed, preferably in terms of content-units and the question may be arranged in increasing order of hierarchy of instructional objectives. Having arranged this, the editing of the paper may be done, which may involve checking each question with respect to the objectives tested, the correctness of the key, the homogeniety of the distractors, the appropriateness of the stem, the preciseness of the language used and the technical format of each question. There are quite a few clues which creep in and care may be taken during editing that these clues are removed. An illustration of improving a multiple-choice item is given in the annexure-A. While editing instructions to the examinees, they may also be looked into carefully. Where-ever two sections, one for the objective-type questions and the other for the free-response are given the specified time limit may also be checked, in terms of the number of questions and their difficulty-level.

7. Preparation of Scoring Key and Marking Seheme

The scoring key and the marking-scheme must be developed simultaneously with the framing of questions. In objective-type questions, the key must be checked and indicated while in case of short-answer questions, the complete answer indicating clearly the credit points, should be made. In case of essay-type

questions, the expected-outline answers, alongwith marks for each value points, should be indicated. Each of these value points must be given weightage proportionately. Where more than one answer is possible and the students have to choose a limited number, the marking scheme should indicate that any of the given points, as expected answer, will be taken as correct. Efforts may be made to minimise the freedom to the examiner to mark in his own way. The best marking scheme is one which when used by many examiners, tends to homogenise the marking pattern.

8. Preparation of the Question-wise Analysis

The last step is the analysis of each question. This can be done by means of proforms, information about each question is inserted under various columns which usually consists of objective specification, forms of questions, unit/sub-unit, marks allotted, estimated time requirement, and estimated difficulty level. From these details, when summarised in a tabulated form (4) (see Annexure-B) one can get an idea of the distribution of questions with different objectives in relation to various units. This helps to validate it against the blue-print developed for the question paper. The analysis can also be undertaken by indicating the serial number of a question at the appropriate place in the blue-print to see if it satisfies the dimension of the blue-print. The question-wise analysis also helps the paper-setter to judge whether the question as a whole would be difficult, easy or is likely to cater to all the three types of students average, poor and bright. Even at this stage, certain questions may be made difficult or easy besides, of course, rectifying the incongruency or inadequacy which might be observed during analysis.

9. Predominance of Objectives in Paper Setting

9.1. *Performance Standard and Objectives*

Once instructional objectives are formulated for a unit of teaching and instruction is done towards the attainment of those objectives, pupils' performance has to be judged in terms of

standards laid in the form of predetermined specific objectives or intended learning outcomes. Students' score on different objectives indicate their level of performance. Same score on two different objectives reflect different standards. A student's score of 20 on Knowledge objective and 20 marks of another student on Application objective is no comparison to make. Performance standard of the latter is likely to be higher than the former. Such an intangible difference which is significant from the point of view of learning, is sub-merged in the ocean of totality of marks. However, for evaluation in school, this anomaly can be obviated by comparing student's performance in terms of the three aforesaid objectives. In view of the increasing importance being given to internal assessment in the new pattern of schooling and the apprehension of lowering of standards, a question paper providing for questions set on all the three levels, can be profitably used to make inter institution and even inter state comparison of students.

9.2. *Why not only Application Questions*

While setting a question paper, it is desirable to give the weightage to knowledge Understanding and Application objectives. This weightage can be worked out in the light of previous standards of student's performance in school or in the Board's examination. An analysis of previous three year's results, preferably objective-wise which could be possibly on a sample basis can form the basis for allocating weightages to different objectives. Gradual decrease in weightage to knowledge objective and the corresponding increase in the other two objectives. Say every fourth year will surely lead to improvement in the standards.

If the three objectives are in hierarchical order as claimed earlier, will it then not suffice to have only application questions which automatically test understanding and knowledge? This should work alright when setting a question paper is meant for inter-class, inter-school or inter-state comparison. But when used for formative evaluation in day-to-day classroom testing, this would not work because the purpose is not merely the

measurement of student's achievement but also improvement of their achievement. For this, the tool must provide evidence which helps to discriminate the students at various ability levels. That is way the questions testing all the three objectives need be included.

9.3. *Content versus Objective*

Content is the medium for appraising the 'students' ability implied by each objective. Should content determine objectives or objective should determine the content? In fact, both are complementary. It is the analysis of content that reveals the possibility of an objective which could be achieved through that content. Every concept may not be amenable to testing at all levels of objectives. On the other hand, if objectives are identified first then a specific content or a concept can be selected to test that objective. In unit testing the former approach may be more suitable while in a full fledged question paper based on the whole syllabus the latter approach is desirable. So complementarity of objective and content need be appreciated.

9.4. *Question Form and Objective*

Instructional objective if clearly understood and predetermined it facilitates the selection of the form of questions to be included in the question paper. In fact the from of question is conditioned by the objective to be tested. For example, if we are interested in testing of abilities like that of investigation, summarisation, expression, reorganisation or evaluation the form of question that suits is essay type only. Like-wise if knowledge of fundamentals of a subject is the aim as in the case of a master test, the objective type questions of multiple choice variety may suit more from the point of view of wider coverage. If one is interested in reviewing lesson in terms of objectives in a limited time then short answer type questions suit more. In class room testing it is the best form for oral testing. Cognizance of objectives is, therefore, essential in deciding about the form of questions to be included in a question paper.

9.5. *Options and Objective*

Place of options in paper setting is quite controversial. Overall option in an external examination is discouraged to make it a more reliable tool of measurement. Internal option of 'this' type has come to stay and their use in internal and external examinations cannot be ruled out. Nevertheless, it is not the option that is important but what the optional questions measure? Unless the optional questions are comparable in terms of what they test for, their validity may be jeopardised. A question testing ability to apply if given against an optional question testing more recall or even understanding, cannot be justified. Optional questions should test the same objective apart from using the same form of questions requiring approximately the same length of expected response and content area. When all these requirements are met is makes the optional question balanced. Commonality of objective is the first criterion when one or more options to the same questions are provided. Ignoring this aspect means sacrificing a lot of validity of the question paper at the alter of options.

9.6. *Objectives and Sections/Parts*

In some question papers there are two or three parts or sections and students are expected to choose a given number of questions e.g. in case of Social Studies where History, Civics, Economics are the major areas of testing. A similar case is sometimes in Biology where Botany and Zoology portions are separated in different sections and students are asked to attempt at least, say two questions from each part out of a total of six. In such cases, there is again every possibility that the choice of question within a section may prompt the students to select simple knowledge based questions in preference to understanding or application based questions which are more difficult to attempt. Here again we sacrifice a lot of validity unless care has been taken to keep all questions comparable in terms of objectives and content areas.

9.7. *Intended Learning Outcomes and Editing*

Analysis of a number of question papers of various Boards of Secondary Education by this writer reveals that even when questions are based on different objectives and stipulated weightages to various objectives is maintained the true nature of an objective is not reflected. For example in testing for understanding it is seldom that questions testing student's ability to interpret, extrapolate or infer are included. Most of the questions are based on classification or citing illustrations. Similarly it is not uncommon to found application questions limited to testing their ability to reason and predict. Rarely a question you would find on testing students ability to formulate hypothesis, suggest procedures to test a given hypothesis or to evaluate a given situation. How does this effect the quality of the question paper? Unless there is adequate sampling of the abilities (intended learning outcomes) implied by an objective purpose of evaluation is defeated. As shown in the diagram depicting relationship of the three objectives, it is the specified objective i.e. the intellectual abilities that indicate the scope of each objective. The more adequately the specifications are covered under each objective the more valid the question paper would be.

Since teaching is an integrated act one is neither expected to teach or each of the specifications one by one, nor it is desirable to do so. Nevertheless, in evaluation vis-a-vis paper setting, deliberate attempt must be made to ensure adequate coverage of these specifications. Some of these can be tested in isolation while others car be integrated in case of questions. In nut shell what is needed is the adequate coverage of the abilities listed under each objective.

9.8. *Objectives and Question-wise Analysis*

Analysis follows the collection of data after administration of the test. If the test results are analysed in terms of objective; as it could be in case of board examinations, or even in terms of specified objectives, as it could be possible in case of classroom testing, it has a great diagnostic value. It provides data about

student's strengths and weaknesses in terms of the developed and the developing abilities. In addition it gives indirectly an indication of instructional impact on the development of those abilities. Besides content wise analysis, the objective of students' achievement is necessary to appraise students' growth in terms of intended learning outcomes, so that instruction if needed be improved.

The same principle applies to reporting and feed back mechanism. If interpretation is done in terms of objectives and evidence is feed back by the Boards to the institutions it can provide the teachers, administrators and in service agencies with useful data to readjust instructional objectives, adapt instructions to the modified goals and affect needed improvement in the measuring devices of evaluation. Such an approach brings in realism in the teaching learning process. On the one hand instructional objectives become the basis of testing, and on the other hand they act as criterion behaviour against which both teaching and testing are validated.

10. To Sum Up

From the above discussion on the technology of paper-setting, it is quite obvious that the farming of a good question paper is not an over-night job but requires a good deal of insight into the basic frame of reference which provides guidelines on the one hand and restrictions on the other, regarding the framing of questions of a particular variety with reference to a particular unit of teaching. If a paper setter is cognizant of the validity and reliability of his question paper, he should follow all these steps listed above. If something goes wrong with the paper-setting, nothing can be set right later regarding judgements made by the evaluators about students' learning.

ANNEXURE A

IMPROVING A MULTIPLE CHOICE ITEM

1. Our body is made up of various types of cells, tissues, organs and the systems. There is some relationship between a tissue and the organ. Same relationship does not exist except between one of the following.
   A. Fats and carbohydrates
   B. Species and genra
   C. Birds and fishes
   D. Plants and animals
   ***E. An organ and system in the body**

2. Our body is made up of various types of cells, tissues, organs and the systems. There is some relationship between a tissue and the organ. Same relationship does not exist between one of the following.
   A. Fats and carbohydrates
   B. Species and genra
   C. Birds and fishes
   D. Plants and animals
   ***E. An organ and system in the body**

   (Double negative in stem is removed)

3. Our body is made up of various types of cells, tissues, organs and systems. There is some relationship between a tissue and the organ. Same relationship exists between one of the following.
   A. Fats and carbohydrates
   B. Species and genra
   C. Birds and fishes
   D. Plants and animals
   ***E. An organ and system in the body**

   (Negative 'does not' is removed in the stem)

4. There is some relationship between a tissue and the organ. Same relationship exists between the following:
   A. Fats and carbohydrates
   B. Species and genra
   C. Birds and fishes
   D. Plants and animals
   ***E. An organ and system in the body**

   (Instructional side is removed in the stem)

5. There is some relationship between a tissue and the organ, the same relationship exists:
   A. Between fats and carbohydrates
   B. Between species and genra
   C. Between birds and fishes
   D. Between plants and animals
   E. Between an organ and system in the body

* is the key expected

(Task is set in the stem)

6. Relationship between a tissue and the organ is the same as
   A. Between fats and carbohydrates
   B. Between species and genra
   C. Between birds and fishes
   D. Between plants and animals
   *E. Between an organ and system in the body

   (Language of stem is made precise)

7. Relationship between a tissue and the organ is similar to the one
   A. Between fats and carbohydrates
   B. Between species and genra
   C. Between birds and fishes
   D. Between plants and animals
   *E. Between an organ and system in the body

   (Language made more appropriate in the stem)

8. Relationship between a tissue and the organ is most similar to the one
   A. Between fats and carbohydrates
   B. Between species and genra
   C. Between birds and fishes
   D. Between plants and animals
   *E. Between an organ and system in the body

   (Language made more specific by adding qualifer in the stem)

9. Relationship between a tissue and the organ is most similar to the one between
   A. Fats and carbohydrates
   B. Species and genra
   C. Birds and fishes
   D. Plants and animals
   *E. An organ and system in the body

   (Stem is made more inclusive to avoid repitition of 'between')

10. Relationship between a tissue and the organ is most similar to the one between
    A. Fat and a carbohydrate
    B. Species and a genus
    C. Bird and a fish
    D. Plant and an animal
    *E. An organ and system in the body

    (Number clue of 'key' is removed)

11. Relationship between a cell and the tissue is most similar to the one between
    A. Fat and carbohydrate
    B Species and genus
    C. Bird and fish
    D. Plant and animal
    *E. An organ and system in the body

    (Verbal clue in stem 'organ' is removed)

12. Relationship between a cell and the tissue is most similar to the one between

A. Fat and a carbohydrate
B. Species and a genus
C. Bird and a fish
D. Plant and an animal
***E. Organ and the system in the body**

(Article clue of 'an' is removed from key)

13. Relationship between a cell and the tissue is most similar to the one between
A. Fat and a carbohydrate
B. Species and a genus
C. Bird and a fish
D. Plant and an animal
***E. Organ and the system**

(Length clue from the key 'E' is removed)

14. Relationship between a cell and the tissue is most similar to the one between
A. Fat and a carbohydrate
B. Host and a parasite
C. Bird and a fish
D. Plant and an animal
***E. Organ and the system**

(Indistinct Distractor—B is replaced)

15. Relationship between a cell and the tissue is most similar to the one between
A. Fat and a carbohydrate
B. Host and a parasite
C. Bird and a fish
D. Liver and the gall bladder
***E. Organ and the system**

(Distractor—D is made more homogeneous)

16. Relationship between a cell and the tissue is most similar to the one between
A. Fat and a carbohydrate
B. Host and a parasite
C. Bird and a reptile
D. Liver and the gall bladder
***E. Organ and the system**

(Inconsistency in Distractor—C is removed)

17. Relationship between a cell and the tissue is most similar to the one between
A. Fat and a carbohydrate
B. Host and a parasite
C. Ectoderm and an endoderm
D. Liver and the gall bladder
***E. Organ and the system**

(Distractor—C is made more plausible in content)

* is the key expected

18. Relationship between a cell and the tissue is most similar to the one between
    A. Fat and a carbohydrate
    B. Ptyalin and the starch
    C. Ectoderm and an endoderm
    D. Liver and the gall bladder
    ***E. Organ and the system**

    (Distractor—B is made more plausible in content)

19. Relationship between a cell and the tissue is most similar to the one between
    A. Ileum and the intestine
    B. Ptyalin and the starch
    C. Ectoderm and the endoderm
    D. Liver and the gall bladder
    ***E. Organ and the system**

    (Distractor—A is made more plausible in content)

20. Relationship between a cell and the tissue is most similar to the one between
    A. Ileum and intestine
    B. Ptyalin and starch
    C. Ectoderm and endoderm
    D. Liver and gall bladder
    ***E. Organ and system**

    (Language of options is made more concise)

21. Relationship between a cell and the tissue is most similar to the one between
    A. Ileum and intestine
    B. Ptyalin and starch
    C. Ectoderm and endoderm
    D. Liver and gall bladder
    ***E. Neurone and the nerve**

    (Key is made more homogenous)

22. Relationship between a cell and the tissue is most similar to the one between
    A. Ileum and the intestine
    B. Ptyalin and the starch
    C. Ectoderm and the endoderm
    D. Liver and the gall bladder
    E. Neurone and the nerve

    (Language of option is made grammatically more appropriate)

23. **Relationship between a cell and the tissue is most similar to the one between**
    A. Ileum and the intestine
    B. Ptyalin and the starch
    C. Ectoderm and the endoderm
    D. Liver and the gall bladder
    E. Neurone and the nerve

    Most (Format of item is improved)

is the key expected

ANNEXURE-B

SAMPLE QUESTION PAPER IN BIOLOGY

Design

QUESTION PAPER/UNIT TEST

Subject : Biology
Unit/Paper : I
Class : IX
Time : 2½ hrs.
Marks : 50

1. Weightage to Objectives:

Objectives	K	U	A	S	Total
Percentage of marks	40	32	16	12	100
Marks	20	16	8	6	50

2. Weightage to Form of Questions:

Forms of Questions	E	SA	VSA	O	Total
No. of Questions	2	12	14	16	44
Percentage of marks	**20**	**50**	**15**	**15**	**100**
Marks Allotted	10	25	7	8	50
Estimated Time	50 mts.	60	14	16	2½ hrs.

3. Weightage to Content:

Units/Sub-Units	Marks
1. Living Organisms	4
2. The Cell	2
3. Cell Wall, Cell Organelles and Cell inclusions	5
4. Cell Division	4
5. Plant and Animal tissues	5
6. Classification	8
7. Bacteria and viruses	3
8. Flowerless plants	6
9. Flowering plants	8
10. Animal Life	6
Total :	50

4. **Scheme of Sections** : Section-A consisting VSA and Objective type questions of 30 mts. duration to be administered separately.

5. **Scheme of Options** : No option is provided.

6. **Difficulty Level** : Difficult : 20% marks.
   Average : 60% marks.
   Easy : 20% marks.

Abbreviations : K (Knowledge), U (Understanding), S (Skill), E (Essay type), SA (Short Answer type), VSA (Very Short Answer type), O (Objective type).

BLUE PRINT

Subject : Biology | Class X

Paper:..... | Maximum Marks 50 | Time 2½ hrs.

Sl. No.	Objectives	Knowledge				Understanding				Application				Skill		Sub-Total	Total
	Units or Sub-Units ↓ / Form of → Questions (1)	E	SA	VSA	O	E	SA	VSA	O	E	SA	VSA	O	E.	SA		
		(2)				(3)				(4)				(5)		(6)	
1.	Living Organisms	—	—	1(2)	½(1)	—	2(1)	—	½(1)	—	—	—	—	—	—	4(5)	4
2.	The Cell	—	—	½(1)	½(1)	—	—	½(1)	½(1)	—	—	—	—	—	—	2(4)	2
3.	Cell wall, Cell Organelles and Cell Inclusions	—	—	1(2)		—	2(1)	—	—	—	—	—	—	—	2(1)	5(4)	5
4.	Cell Division	—	2(1)	—	½(1)	—	—	½(1)	½(1)	—	—	—	½(1)	—	—	4(5)	4
5.	Plant and animal tissues	—	—	1(2)	1(2)	—	—	—	½(1)	—	2(1)	½(1)	—	—	—	5(7)	5
6.	Classification	5(1)	—	—	—	—	2(1)	—	½(1)	—	—	—	½(1)	—	—	8(4)	8
7.	Bacteria & Viruses	—	2(1)	—	—	—	—	½(1)	—	—	—	½(1)	—	—	—	3(3)	3

(1)	(2)				(3)				(4)				(5)	(6)	
8. Powerless plants	—	—	—	—	—	2(1)	½(1)	½(1)	—	—	—	—	—	3(1)	6(4)	6
9. Flowering plants	4(1)	—	—	½(1)	—	—	—	½(1)	—	2(1)	—	—	—	1(–)	8(4)	8
10. Animal Life	—	—	½(1)	—	—	2(1)	—	½(1)	—	2(1)	—	—	—	—	5(4)	5
Sub-Totals	9(2)	4(2)	4(8)	3(6)	—	10(5)	2(4)	4(8)	—	6(3)	1(2)	1(2)	—	6(2)	50(44)	50
Totals	20				16				8				6		50	

Notes : Figures within brackets indicate the number of questions and figures outside the brackets indicate marks.

*Denotes that marks have been combined to form one question.

Summary : Essay (E)	No. 2	Marks 10	Scheme of Options.....
Short Answer (SA)	No. 12	Marks 25	Scheme of Sections.....
Very Short Answer (VSA)	No. 14	Marks 07	
Objective (O)	No. 16	Marks 08	

APPENDIX

SAMPLE QUESTION PAPER BIOLOGY

Marks : 50
Time : 2½ hrs.

Instructions

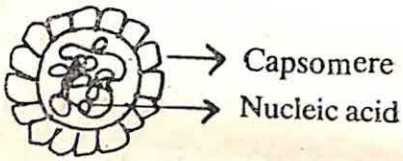
1. All questions are compulsory and carry half mark each.
2. In Section-A, Question Nos. 1-14 require very short answer which may be one word, one phrase, or a sentence. Answers to these questions are to be written on the space provided in the question paper itself.
3. In questions 15 to 30 only one answer out of the 4 given is correct. Serial number of the correct answer is to be encircled.
4. Section-A will be collected back by the invigilator after 30 minutes.
5. In Section-B, marks are given against such question from Q. No. 31 to 44.
6. Questions 31 to 42 require short answer involving 2 to 4 credit points and can be answered in about 20 to 40 words. Only precise answers are expected.
7. Question Nos. 43 and 44 require long answers which should not exceed 400 words or so.

Section—A

Time : 30 mts.

1. What term is used to the characteristic of living organisms to respond to alteration in environment?
2. What term is used for the folds of the inner membrane of mitochondria projecting towards its interior?
3. If a child is gradually loosing weight which of the two processes Anabolic or Catabolic may be faster?
4. Which cell organelle is called the suicidal bag of the cell?
5. Which biologist postulated for the first time that every cell arises from the pre-existing cell?

6. Give an example of a cell which changes its shape very frequently.
7. Corresponding genes have the same set of trait. What is such a pair called?
8. Which fundamental tissue in plants helps to prevent tearing of plant body when strong wind is blowing?
9. Which of the two phases, the sporophytic or gametophytic is dominant in thallphytes?
10. In which of the two bacterial cells, having the same number of mesosomes but different surface area—volume ratio rate of metabolism will be more?
11. Observe the diagram given opposite and identify the organism.



12. Which organelle in Amoeba serves the same function as the kidney cell in an animal?
13. Small terminal air sacs of the lungs concerned with gaseous exchange are called .....
14. What is the basic unit of a nervous tissue?
15. Which of the following organisms can be placed both at cellular and individual level of biological organism?

    A. Liver
    B. Amoeba
    C. Rhizopus
    D. Hydra

16. Cell theory does NOT state that all cells

    A. perform metabolic activity.
    B. arise from pre-existing cells.
    C. contain hereditary material.
    D. originate spontaneously.

17. A pair of two longest cells are:

    A. Amoeba and nerve cell.
    B. Ostrich egg and human cell.

C. Nerve cell and cotton fibre cell.
D. Muscle cell and kidney cell.

18. Number of chromosomes in the body cell of man, dog and horse are in the increasing order of
A. man—horse—dog.
B. dog—horse—man.
C. man—dog—horse.
D. horse—dog—man.

19. Which of the following illustrations involves meiotic cell division?
A. Healing of wounds.
B. Regeneration of damaged parts.
C. Formation of tumours.
D. Source of new genetic variations.

20. In which of the following phases of Mitosis centromere of each Chromosome splits into two Chromatids?
A. Prophase
B. Mataphase
C. Anaphase
D. Telophase

21. As a result of injury internal tissues comes in direct contact with
A. Columnar epithelium.
B. Cuboidal epithelium.
C. Squamous epithelium.
D. All the above.

22. At maturity which of the following is the living component of Xylem?
A. Xylem fibres.
B. Xylem parenchyma.
C. Xylem tracheids.
D. Xylem vessel.

23. Which of the following is NOT a fish?
A. Sucker fish.
B. Flying fish.
C. Star fish.
D. Dog fish.

24. Which of the following characteristics can help you to identify a pteridophyte from bryophyte?

A. Dominant gametophyte.
B. Sporophyte dependant on gametophyte.
C. Gametophyte dependant on sporophyte.
D. Dominant sporophyte.

25. Which of the following is a modification of stem but serves the function of a leaf?
A. Tendual of sweat pea.
B. Phylloclade of opuntia.
C. Pitcher of Nepenthes.
D. Phyllode of Acacia.

26. A fish when taken out of water immediately dies because
A. gill slits close in the absence of water.
B. gill slits open too much in air.
C. respiration through skin does not take place.
D. fish is a cold blooded animal.

27. In which of the following animals blood play no part in respiration?
A. Cockroach.
B. Earthworm.
C. Frog.
D. Fish.

28. Which of the following structures is concerned with asexual reproduction?
A. Egg.
B. Sperm.
C. Bud.
D. Gamete.

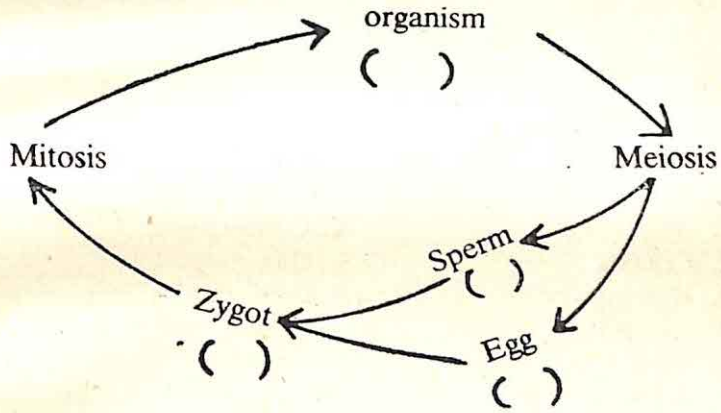
29. In which of the following tissues columnar expithelium is found?
A. Alveoli of lung.
B. Epidermis of skin.
C. Germinal layer of ovary.
D. Lining of intestine.

30. Naked seeds are found in both
A. Pinus and cedrus.
B. Maize and wheat.
C. Date palm and pinus.
D. Cedrus and bean.

Section—B

Time : 2 hrs.

31. Give one distinguishing feature of temporary and one of permanent adaptation citing one example of each. 2
32. In what way a plant cell differs or resembles an animal cell with respect to organelles concerned with respiration, food manufacture, storage of waste products and protective functions? 2
33. Draw an internal structure of chloroplast and label granum, stroma, lipid droplet and starch grain. 2
34. In the given sketch fill up haploid or diploid number (n/2n) of chromosomes in the brackets provided.



35. What role blood corpuseles play in formation of pus in a wound?
36. Group the following plants into two groups giving the basis of your classification.
    (a) Funaria (b) Pinus (c) Selaginella
    (d) Rhizopus (e) Spirogyra (f) Cedrus. 2
37. Name two edible fungi and two having medicinal use. 2
38. List four applications of uses of bacteria. 2
39. Draw a vertical section of thallus of merchantia showing rhizoids, storage region, photo-synthetic region, upper and lower epidermis with one air pore. 3
40. What would be the effect on the osmo-regulation func-

tion of the organism when

(a) Marine species of amoeba is transferred to fresh water?

(b) Fresh water species is transferred to sea water? 2

41. A plant has well developed root system, succulent stem leaves modified into spines. What could be the habitat of such a plant and why? 2

42. Skeleton of sharks which live at great depths is made up of cartilage only. What difference would it have made if they had bony skeleton? 2

43. Develop a flow chart depicting classification of the phylum chordata into five main sub-phyla. List two unique characteristics of this phylum and compare the characteristics of fishes and amphibians. 5

44. Draw an outline diagram showing the transverse section of a dicot root and describe the structure and function of various tissues. 5

Key—Outline Answers and Marking Scheme

Very Short Answer

1. Irritability
2. Cristae
3. Ketabolism
4. Lysomes
5. Rudolf Virchow
6. Amoeba/w.Bc
7. Homologous pair
8. Collenchyma
9. Gametophytic
10. Cell with higher ratio of surface area to volume
11. Virus
12. Contractile vacuole
13. Alveoli
14. Neurone/nerve cell

Multiple Choice

15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
B	D	C	A	D	C	C	B	C	D	B	A	A	C	D	A

31. (a) In temporary adaptation individual adapts to some specific conditions in environment/ cannot be inherited.

(b) For example bending of seedling towards whole when kept in a dark box with a whole.

(c) In permanent adaptation species exhibit permanent changes over several germination/ genetic basis.

(d) For example fish is adapted to water and opuntia to desert conditions. 2

32. (a) Presence of cell wall for protection of plant cell.

(b) Presence of chlowplast for food manufacture in plant cell.

(c) Presence of mitochondria in both.

(d) Presence of vacuole for storage of waste product in plant cell. OR any other. 2

33. Correct labelling and drawing of

(a) granum.

(b) stroma.

(c) lipid droplet.

(d) starch grain. 2

34. (a) Organism—2n

(b) Sperm—n

(c) Egg – n

(d) Zygot—2n $4 \times \frac{1}{2}$ 2

35. (a) W.B.C. are fighting cells of blood.

(b) Bacteria attach the cells in the wound.

(c) W.B.C. of the blood fight with bacterial cells.

(d) Pus is formed by the ramnants of the killed bacterial cells and W.B.C. $4 \times \frac{1}{2}$ 2

36. (a) Funnaria, Selaginella, rhizopus and spirogyra.

(b) Cedrus and pinus.

(c) Basis for (a) is flowerless/seedless plants.

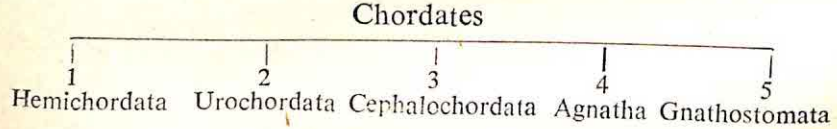
(d) Basis for (b) is flowering/seed plant. $4 \times \frac{1}{2}$ 2

37. Any two like:

(a) Edible—Morchella and Agorious. $2 \times \frac{1}{2}$ 1

(b) Medicinal—Pencillium and Gibberella. $2 \times \frac{1}{2}$ 1

38. Any four like:
    (a) Tanning of leather.
    (b) Fermentation of alcohol.
    (c) Curdling of milk.
    (d) Sewage disposal by decomposition.
    (e) Fixation of atmospheric nitrogen. $4 \times \frac{1}{2}$ 2
39. Neat and correct diagram showing correct location of
    (a) Rhizoids.
    (b) Storage region.
    (c) Photosynthetic region.
    (d) Upper epidermis.
    (e) Lower epidermis.
    (f) Ear spaces. $6 \times \frac{1}{2}$ 3
40. (a) Need for elementation of excess water absorbed arises.
    (b) Development of Contractile vacuoles.
    (c) Need for elimination of excess water stops.
    (d) Contractile vacuoles disappear. $4 \times \frac{1}{2}$ 2
41. (a) The plant could be a Xerophyte.
    (b) Well developed root system shows lack of water available.
    (c) Succulent stem shows the need for storage of water.
    (d) Modification of leaves into spines shows the need for conservation of water. $4 \times \frac{1}{2}$ 2
42. (a) Cartilage is flexible.
    (b) Bone unlike cartilage, is not flexible.
    (c) It will not allow fish to endure pressure.
    (d) Thus it will be fatal to the fish. $4 \times \frac{1}{2}$ 2
43. Flow chart like the following is expected.



    (a) If all sub-phylla are correct.
    (b) Three characteristics: $3 \times \frac{1}{2}$ $1\frac{1}{2}$

(i) Hollow dorsal tubular nerve cord.
(ii) Noto chord.
(iii) Gill in the pharynx.

(c) Fishes and amphibian: description of any five: $5 \times \frac{1}{2}$ $2\frac{1}{2}$

(i) Both are cold blooded.
(ii) Presence of jaw in both.
(iii) Skin in both is slimy but devoid of scales in amphibians.
(iv) Respiration through gills in fishes but the amphibians it is through skin, gills or lungs.
(v) Heart of fishes has one auricle and one ventricle but it has two auricles and one ventricle in amphibians.

44. (a) Outline diagram. 1 5

(b) Description of structure and function of

(i) Epidermis.
(ii) Cortex.
(iii) Endodermis and Pericycle.
(iv) Vascular bundle. 4×1 5

QUESTION-WISE ANALYSIS

Sl. No.	Objective	Specification	Content Unit/Sub-Unit	Form of Question	Marks Allotted	Estimated Time	Estimated Difficulty level
1	2	3	4	5	6	7	8
1.	K	Recalls	Unit-1	VSA	½	1 mt.	C
2.	K	Recalls	Unit-3	VSA	½	1 mt.	B
3.	K	Recalls	Unit-1	VSA	½	1 mt.	B
4.	K	Recalls	Unit-3	VSA	½	1 mt.	B
5.	K	Recalls	Unit-2	VSA	½	1 mt.	C
6.	K	Recalls	Unit-2	VSA	½	1 mt.	B
7.	U	Translates	Unit-2	VSA	½	1 mt.	B
8.	A	Analyses	Unit-5	VSA	½	1 mt.	A
9.	U	Compares	Unit-8	VSA	½	1 mt.	B
10.	A	Infers	Unit-7	VSA	½	1 mt.	A
11.	U	Interprets	Unit-7	VSA	½	1 mt.	B
12.	U	Translates	Unit-12	VSA	½	1 mt.	B
13.	K	Recalls	Unit-5	VSA	½	1 mt.	C
14.	K	Recalls	Unit-5	VSA	½	1 mt.	C
15.	K	Recognises	Unit-1	MC	½	1 mt.	C
16.	K	Recognises	Unit-2	MC	½	1 mt.	B
17.	U	Cites examples	Unit-2	MC	½	1 mt.	A

18. K	Recognises	Unit-4	MC	½	1 mt.	B
19. A	Analyses	Unit-4	MC	½	1 mt.	B
20. U	Compares	Unit-4	MC	½	1 mt.	B
21. U	Interprets	Unit-5	MC	½	1 mt.	A
22. K	Recognises	Unit-5	MC	½	1 mt.	C
23. U	Compares	Unit-6	MC	½	1 mt.	B
24. U	Compares	Unit-8	MC	½	1 mt.	A
25. U	Interprets	Unit-9	MC	½	1 mt.	C
26. A	Reasons	Unit-6	MC	½	1 mt.	A
27. U	Interprets	Unit-10	MC	½	1 mt.	A
28. U	Relates	Unit-1	MC	½	1 mt.	C
29. K	Recognises	Unit-5	MC	½	1 mt.	C
30. K	Recognises	Unit-7	MC	½	1 mt.	C
31. U	Compares	Unit-1	SA	2	5 mts.	C
32. U	Relates	Unit-3	SA	2	5 mts.	B
33. SK	Draws	Unit-3	SA	2	5 mts.	A
34. K	Recalls	Unit-4	SA	2	5 mts.	A
35. A	Analyses	Unit-5	SA	2	5 mts.	B
36. U	Classifies	Unit-6	SA	2	5 mts.	A
37. U	Cites examples	Unit-8	SA	2	5 mts.	B
38. K	Recalls	Unit-7	SA	2	5 mts.	C
39. SK	Draws	Unit-8	SA	2	5 mts.	B
40. A	Predicts	Unit-10	SA	2	5 mts.	A
41. U	Explains	Unit-10	SA	2	5 mts.	B

1	2	3	4	5	6	7	8
42.	A	Predicts	Unit-10	SA	2	5 mts.	A
43.	K	Recalls	Unit-6	E	5	25 mts.	B
44.	K	Recalls	Unit-9	E	5	25 mts.	B

Note : Use abbreviations as follows:

(a) For Objectives: K=Knowledge, U=Understanding, A=Application, S=Skill.

(b) For form of questions: O=Objective, VSA=Very short-anwer, SA=Short-answer, E=Essay.

(c) For difficulty level of questions: A=Difficult, B=Average, C=Essay.

CHAPTER XV

MODERATION OF A QUESTION PAPER

1. Pre-examination Moderation

The term moderation is used with different connotations. Moderation of questions and question papers, moderation of marks, moderation of results and even moderation of syllabi. In this chapter, however, we restrict ourselves to moderation of a question paper and marks. Moderation means re-examination of the questions, marking scheme, key, model answers (if given) and instructions/directions for examinees besides format and mechanics of paper setting. Purpose is to ensure quality, relevance, reliability and practicability of the question paper in a given situation, i.e. home examination or a public examination like the Boards' examination. This exercise is done with the purpose of improving the question paper in terms of requirements reflected in the given design and the blue print of the question paper.

After going through this chapter the reader should be able to:

(a) recognise the need for moderation of question papers and marks.
(b) identify the various tasks implied in the moderation of a question paper.
(c) apply various steps involved in post-examination moderation through expert judgement.

(d) use statistical methods for adjusting marks awarded by different teachers or evaluators.
(e) recognise the limitations of various moderation procedures in improving the quality of question papers, vis-a-vis reliability of scores.

Moderation in case of internal examination can be undertaken by paper setter himself or more appropriately by a co-teacher of the same subject teaching the same class or the higher classes. In case of external examinations conducted by the boards, the moderator is always a different person from the one who sets the paper. It is also assumed that the moderator is a subject expert of long standing, well conversant with the technology of paper setting and is quite cognizant of the instructional objectives which form the basis of framing questions. With a knowledgeable moderator quality of a question paper improves but its quality may go down if the moderator is not well conversant with role of instructional objectives in paper setting and the technique of paper setting.

2. Tasks Implied in Moderation of a Question Paper

Since purpose of moderation is to improve the quality of question paper the process of moderation includes a number of tasks which a moderator has to undertake with respect to different aspects of a question paper.

(a) Study of the prescribed design for understanding the basic framework.
(b) Verification of blue print developed by the paper setter, to judge the inconsistency if any, with regard to the prescribed design.
(c) Editing of each question in terms of blue print, using known criteria of framing a question and making improvements if any. Special care has to be taken to the use of directional words like discuss, comment, justify, what do you know of etc. etc. which are often misused.
(d) Checking of key, outline answers and marking scheme to judge the correctness of model answers/key in terms of

intended response for a question and scoring objectivity.

(e) Judging appropriateness of organisational pattern of consolidating questions in the form of sections or question paper as a whole, from the point of view of ease of administration scoring and interpretation.

(f) Review of general instructions to judge their relevance, adequacy of total time and section-wise allotment of time, (as in case of objective type questions).

(g) Question-wise analysis of the whole question paper to see whether, objective tested, topic covered, form of questions used, marks allotted, difficulty level estimated, and estimated time for each question tally with the blue print requirement vis-a-vis the design prescribed.

(h) Make necessary modifications in questions, instructions and format of the question paper in the light of observations made in (a-h) above.

(i) Read through the final version once again to check for language, spellings, serialling of questions and other aspects of mechanics of writing, to finalise the paper.

3. Some Dos' and Donts'

(a) A moderator has no right to modify the design prescribed by the examining agency.

(b) A moderator need not alter the blue print unless there is some inconsistency in distribution of questions and marks relating to objectives or topics. In that case he is supposed to make corresponding changes in blue print or in the distribution of questions.

(c) A moderator has no right to change weightages given to various objectives, units or topics and form of questions reflected in the blue print which in turn conforms to the requirements of the design.

(d) A moderator has no right to change the form of question but he is free to improve, modify, change or replace a question to suit the blue print.

(e) A moderator must estimate the time requirement of each question as also for the whole paper to judge the adequacy or inadequacy of the allotted time. If separate time limit is

prescribed for different sections, its appropriateness in terms of time requirement should also be estimated and adjusted.

(f) A moderator is completely free to modify the key, model answer or marking scheme to dove-tail it to the intended answer of the question.

(g) In case of free options i.e. any 5 out of 8 questions type, all questions should be judged in terms of difficulty level and time requirement to ensure that they all are almost equivalent in these two aspects.

(h) In case of internal options of Either-Or type the moderator should ensure that both questions or options are equivalent in terms of objective tested, subject area, difficulty level, form of question and the time requirement.

(i) A moderator must ensure that language used in framing question is precise, unambiguous and comprehensible.

(j) A moderator should see to it that format and organisational pattern of question paper facilitate its printing administration and scoring of questions.

(k) He must review the instructions given for candidates to judge for their relevance, adequacy and preciseness.

4. Post-examination Moderator

When question papers, marking scheme and other aspects are finalised and approved the examinees take the test. Their scripts are marked and scores are given. At this stage also moderation of marks is useful. This is not necessary in case of school examinations except in cases where a large number of students are there in a class consisting of many sections and individual teachers assess their wards independently, in annual examination say, for promotion purpose. Of course for public examinations moderation of marks of different evaluators is essential to secure more reliable grading. Two approaches may be used for this.

4.1. *Expert Judgement*

This requires taking advantage of experienced teachers to use

their judgement to adjust marks of different evaluators on the basis of 'mental image' of the intended standard (5). This moderation is based on *overall* standard used by an individual teachers and NOT the actual performance of individual candidates. It involves inspection by experts of candidates' work and comparing each teacher's examinees with the 'mental standard' visualised by the moderator who is considered to be an experienced judge with integrity. Training of moderators, conducting trials for ensuring agreement and arriving at consensus is the usual process followed in Mode-3 examination of UK (6). This is quite applicable to Indian situation with respect to one examining board or more than one examination boards within a state.

Following steps are envisaged for a subject in this type of moderation:

(a) A chief moderator/chief coordinator or paper setter himself is designated by the examining agency to supervise the whole operation. Let us call him a chief moderator.
(b) Chief moderator may invite a representative in the subject from each of the area, called a head examiner/additional head examiner/coordinator etc. Let us call him a coordinator under whom a number of examiners/markers/ evaluators work. Let us call them markers of scripts.
(c) Each coordinator is given an *identical set* (same scripts xeroxed) along with marking scheme/key/outline answers. They are given the specific guidelines and principles underlying marking or grading by the chief moderator.
(d) Each marker grades independently the sample of scripts given to him.
(e) Discussion among markers follows to arrive at consensus over the agreement on the marking scheme on the one hand and similarity of *overall* standards of various markers on the other. Consensus grade will be the average of grades given by different markers.
(f Individual's assessments are then compared with the consensus of the group on the basis of

(a) Overall accordance with the consensus arrived at.

(b) spread of grades among the candidates assessed.

(c) general leniency or severity observed in marking.

4.2. *Leniency versus Severity in Marking*

Just to illustrate if we take the six students i.e. in place of 10 or 20 as suggested earlier for sample marking we may have three types of situations.

Candidate	Consensus grade	Markers' grade		
		Marker-1	Marker-2	Marker-3
1. Amina	3	4	5	2
2. Karina	1	1	2	1
3. Marina	5	4	6	4
4. Rabina	4	4	5	3
5. Sabina	6	5	6	5
6. Zarina	2	1	3	2
Total	21	19	27	17

Highest grade is 1. Half grade per candidate=6/2=3

Marker-1

Total (19) falls within ± 3 of the consensus total of (21) and can, therefore, be considered acceptable overall standard of this marker.

Marker-2

Total (27) does not fall within ± 3 of concensus grade of (21) and, therefore, his overall standard is not acceptable. He errs a lot towards severity and may be considered too severe.

Marker-3

Total of (17) again does not fall within ± 3 of consensus grade of (21). He errs slightly towards leniency. Had his total been (18) his standard would have come within ± 3 of the consensus grade (21) and become acceptable.

Further discussion of criteria and marking scheme would enable the chief moderator to arrive at agreement. It may need sometime one or two additional agreement trials to bring too lenient or too severe examiners/markers within an acceptable range of consensus grade.

Spread of Grades of Candidates

Let us consider the following example involving 6 candidates.

Spread of Grades

Candidate	Consensus Grade	Marker's Grade
1. Bharat	2	1
2. Dasrath	4	3
3. Janak	1	4
4. Laxman	3	4
5. Ram	6	6
6. Sita	5	4
Total	21	22

Here marker's range of grade (1-6) is the same as the consensus grade (1-6) and, therefore, passes the '*spread of grades*' test. He also passes the leniency/severity test as the total of (22) is within the accepted range of (6/2=3.00) consensus grade total (21) i.e. ± 3 of 21. Inspite of that the marker has not used the grades (1-6) similar to the consensus grade (1-6). For example 2 and 5 are not used and there is bunching of three candidates (4, 4, 4) into one grade. Thus range itself is not sufficient to indicate the acceptability of marking/grades. This requires another check for confirmation.

4.3. *Check for Conformity*

The main purpose of this check is to compare each marker's each grade with the consensus grade. Following example illustrates the principle underlying.

Conformity Check

Candidate	Marker's grade (M)	Consensus grade (C)	Difference between grades (M−C)
1. Anny	5	5	0
2. Banny	3	4	−1
3. Canny	1	1	0
4. Danny	2	1	+1
5. Nanny	3	2	+1
		Total =	+1

In this case difference is +1 for five candidates and it is normally accepted for a sample of 20 candidates. Maximum acceptable difference could be 12 and should not exceed that. Certain tables with different samples of various sizes are given by Yorkshire Board (1972) for the three tests mentioned above.

On similar lines it is possible to moderate internal assessments of different schools with the help of Inter-school moderators. What is important is that, the tests viz. 'leniency or severity test', 'spread of grade test' and the 'conformity check' if applied would go a long way in moderating the grades to ensure maximum comparability viz-a-viz reliability. Sometime 'balanced pairing' is suggested by associating a lenient scorer with the severe scorer. The assessment of scripts by another marker is also done sometimes to ensure more objectivity in scoring. Some boards in UK use 'criteria of quality' prescribed for grading minimum requirement for a grade. For example grade-1 could be the evidence of reasoning and for grade-4 say, evidence of recall of facts and events may suffice.

All the above mentioned measures for moderation are based on logic, opinion, experience, judgement and agreement through discussion among the makers, moderators and others who are involved in processing. No specific statistical measures are recommended in arriving at consensus grade.

4.4. *Moderation by Statistical Means*

Statistical methods can also be used for adjusting of marks

awarded by different teachers or evaluators. This can be done by following methods.

4.4.1. *By Use of Moderating Test*

Since teacher's assessments are to be judged on the basis of 'conformity' with the results on the moderating test besides the 'severity' and 'spread' then the reference test (moderating test) must be the one on which candidates' performance could be expected similar to the teachers assessment being moderated. Moderating test must be valid, reliable and discriminating so that candidates performance viz-a-viz ranking on both the tests i.e. teachers' assessment and moderating test, should remain similar. Therefore, moderating test should be identical to actual assessment. A *parallel test* may be constructed by the board or concerned agency and it can be used as reference test for fixing standard of the school. Another way is to use a *common core test* based on the prescribed instructional objectives. Performance on this common test, of candidates from different institutions can help to judge the validity of teachers' assessments without moderating the marks. Of course this is based on the assumption that the different schools are following the same syllabus.

4.5. *General Ability Tests*

Another possibility is the use of content free tests testing abilities like reasoning, basic mathematical skills, etc. which are independent of the syllabi followed. Such tests give the idea of general mental ability of candidates from various schools and this enables the moderator to interpret students' performance on the basis of general ability test. If there is significant difference in the general mental ability scores of student's from different schools the actual subject-wise score on teachers' assessments have to be interpreted in terms of scores on general mental ability tests.

4.6. *Moderation by Standardisation*

Mean and standard deviation of marks awarded by teachers

and marks awarded on the moderating test are computed. Then marks of each student are standardised to the mean and standard deviation of the school results on the moderating test using the following formula:

$$\frac{M_m - X_m}{s_m} = \frac{M_t - X_t}{S_t}$$

Where : M_m is the mean of marks on moderating test (say = 50)

M_t is the mean of marks on teachers' test (say=60)

S_t is the standard deviation of marks on teacher's test (say 5.0)

S_t is the standard deviation of marks on moderating test (say 7.0)

X_t is the marks of a candidate on teacher's test (say 40)

X_m is the moderated marks of that candidate.

This formula is given in many texts like the one by Guilford (7). It is obvious that teacher was lenient (mean score-60) compared to mean of moderating test (Mean score-50). Spread of scores (5.0) is also not as large as of Moderating test (7.0). In the example we can find out the moderated mark as under:

$$\frac{60 - X_m}{7} = \frac{50 - 40}{5} = \frac{10}{5} = 2.0$$

$$\frac{60 - X_m}{7} = 2 \quad \text{OR} \quad 60 - X_m = 14$$

$$X_m = 60 - 14 = 48$$

Thus 60 marks of this candidate awarded by the teacher become 48 moderated marks which are below 60 just as mean score (50) of moderating test is less than that of mean score (60) of teacher's test. This method is useful when a large number of students are appearing as in an examination Board and their marks are moderated.

The above mentioned methods of moderation of marks are useful for comparison of candidates' performance on different tests used by teachers or evaluators. Nevertheless, for classroom testing by teachers one may improve to a great extent

the scoring objectivity and comparability if due precautions are taken at the level of designing the question paper, blue printing, construction of questions, developing marking schemes and during editing of questions and their model answers. Occasional use of common tests in schools can help in surveying the level of achievement of students. Feedback of performance on such tests helps the teachers to make up deficiencies through enrichment or remedial programme which in turn helps in minimising the gap in achievement scores of students from different institutions. Performance on this common paper, of candidates from different institutions can thus help to judge the validity of teachers' assessments without moderating the marks.

Moderation does improve the quality of a question paper and the accompanying marking scheme but only to the extent the quality of the question paper permits. A poorly developed question paper sets the limits to its improvement unless the whole question paper is reconstructed. Like-wise a poorly developed blue print sets the limits for moderation and acts as a limiting factor beyond a particular stage of improvement. Thus process of moderation is conditioned by the limits imposed by the potentialities of the paper built in by the paper setter.

CHAPTER XVI

ANALYSIS FOR IMPROVING QUESTIONS AND QUESTION PAPERS

1. Introduction

It is an erroneous belief that all is going well with the paper setting at the public examinations. In fact the reality is that at least 50% of our examinations have at least 5% error of measurement. This is what the results of Indian examinations indicate. But how many of us as paper setters know this? Despite our best efforts to prepare designs, blue prints and construct-objective-based questions, a paper set by a trained and experienced subject-specialist may yield unintended results. What happens is that the various assumptions that are made during the setting of a question paper may be quite relevant to certain institutions and less relevant to others. Assumption about students' inputs, instructional inputs and mode of evaluation might have varied from place to place. As such, the assumed pre-requisite learning and skills developed during the course of studies may not have actually developed and the result may not be what was expected. Therefore, to know how really the paper generated various types of responses from the students, it is significant to judge the quality of the paper set. It is through proper analysis that we can find out the areas of adequacies and inadequacies in students' learning.

It is expected that after going through this chapter the reader will be able to

(a) recognise the need for analysis of tests and test results.

(b) differentiate between the psychometric and psychiatric (illuminative or descriptive) approaches.
(c) compute reliability of tests using different methods.
(d) calculate facility index and discrimination index of items using different formulae for item analysis.
(e) interpret test results in terms of prescribed criteria objectives.
(f) use item analysis and test results for feedback and improvement of tests and students' learning.

2. Why to Analyse?

It does not suffice to say that a poor or good performance in the test indicates the quality of students' learning. This is based on the assumption that the instrument of evaluation i.e. the question paper had the desired qualities of validity and reliability. In case the question paper lacks the minimum requirement of measurement qualities all efforts that go into the analysis and interpretation of test results are fruitless. Therefore, the first attempt should be to find out the validity and reliability of the testing instrument and then if we find it adequate we can proceed to analyse the test results.

At the Board level, a public examination is not only conducted for grading the students but also for maintaining and improving standards. It is, therefore, essential that the boards should know about the performance of students in general and subject-wise in particular. Unless this is done it is not possible to identify the strengths and weaknesses of students which alone can help the boards to improve evaluation procedures vis-a-vis instruction, through proper feedback of results to the institutions. In a way it helps to the board in self-evaluation, to see the extent of congruence or discrepancy between the intended and the actual outcomes. What are the instructional objectives (as laid down in the prescribed curriculum) which have been achieved and what are the ones which need more attention at the hands of teachers. Therefore, data may be useful for affecting curricular changes vis-a-vis improvement in curriculum.

When results are analysed in terms of subjects, it provides good evaluation of teachers. The poor performance of students

in a particular subject reflect the instructional inputs. Knowledge of subject-wise performance at the state level and institution-wise is necessary if one is interested in teachers' evaluation. Indeed, heads of institutions do judge a teacher on the basis of results in their subjects. But they do not get a comparative view. If proper proformas are developed for examiners' reports, area-wise performance within a subject is also possible and that would be useful for the teachers and curriculum developers to redefine and improve the curriculum. Thus, to my mind four major purposes of analysis are:

(a) to judge the efficiency of the test.
(b) to judge teacher effectiveness.
(c) to provide feedback to institutions to improve teaching learning strategies.
(d) to improve the curriculum and instruments of evaluation.

Except for the second purpose which is realisable better at the school level, the other three are more relevant at the Board level.

3. When to Analyse?

Analysis begins in fact at the time of designing the question paper. Deciding about weightages to the various instructional objectives, content units and form of questions in itself involves identification of significance and preferred objectives, content units and form of questions. This is generally based on previous experience, results or tradition. Anyway, the blue printing of a question paper further involves analysis of content on the basis of which questions are distributed over the various content units to test different objectives. Likewise, the question-wise analysis that is undertaken towards the end of paper setting, is another step towards analysis. It is at this stage that the balance or imbalance of a question paper is judged and modifications if any, needed at this stage, are made. All types of analyses which are done before the administration of a question paper are based on the logic, subjective opinion of the setter or the

judgements of a panel involved at one or the other stage. This sort of judgemental validity is necessary but not a sufficient condition for ensuring the needed validity or reliability, as it is based only on experts' opinion not on any empirical evidence. It is for this reason that we need to know the empirical evidences on the basis of which sound judgements could be formed.

The time for analysis of test results vis-a-vis the question paper is after administration of the paper. When the scripts are marked and the results are prepared, the board can undertake various types of analyses depending upon the purpose and the use to which those results are to be put. This work can be facilitated by developing a structured proforma for the markers (examiners) who may be asked to collate some information about the scripts they mark, in a manner indicated in that proforma. Another time for analyses is after the declaration of the results. This is normally done to study the nature of errors committed by the students in various subjects. Likewise analysis is also possible at a later stage as a part of allied research projects in which data from the same year or previous years of the same board or other boards of school education can be used for appraisal of comparative performance of students. Nevertheless, the minimum that is expected from a board is to get the results analysed subject-wise every year, at least to see how a question paper behaves and how the students received the question paper.

4. Who Can Analyse

Depending upon the purpose and timing of analysis the teachers, the examiners, paper setters, board officials, researchers and even students can undertake analysis. Students reflect much better on the choice of questions made by the paper setter as well as the area covered (content-validity). Teachers' can reflect better on the results in their respective subjects. Examiners are the most suitable judges for reflecting on the quality of responses of students and the probable reasons thereof, besides giving their evidence-based opinion on the vagaries of questions or their potentialities. It is they who can identify the magnitude and nature of errors committed by

students. It is they alone who can opine on the types of errors students make and the errors which can be traced to defective questions or to instructional deficiencies. Paper setters can self-evaluate for adequacies or inadequacies in their paper setting, on the basis of analysis and reports received by the markers (examiners).

Board officials can undertake analysis of results in terms of identification of areas of curriculum improvement as well as the quality of measuring instruments. Researchers can go deeper to analyse the causes of failures, underachievement, low performance and behaviour of different items or questions. They can even probe deep to judge the suitability of various concepts and their placement in the curriculum. Thus, depending upon the interest and administrative decisions, different types of persons can be commissioned for analysis of results.

5. Which Approach to be Used?

There are basically two approaches which can be used. One is the traditional, classical psychometric approach which emphasises precision and prediction and is rooted in the belief that results can be analysed and interpreted in quantitative terms irrespective of the context or antecedents that might have influenced the students' performance. Use of statistical techniques in processing and interpretation of the results is the usual mode of analysis. This is the approach which suits analysis of Boards results where the purpose is mainly to measure students' achievement, classify and grade students rather than improve their achievement. Therefore, the judgements that are to be made are norm-referenced where students' performance is judged in terms of deviation from the norm. Students are declared as pass, fail, underachievers or deviants. Judgements in terms of predetermined criteria or intended outcomes in the form of instructional objectives is relegated to the background.

The other approach is the descriptive, psychiatric, narrative or illuminative approach which is embedded in the social-anthropological paradigm and emphasises the description and interpretation of results for the purpose of improvement of students' learning. The focus is more on diagnosis, narration of

evidences and it takes into account the antecedents, inputs and processes which effect students' learning and achievement. The interpretations are made more in terms of the student himself and with reference to the criteria or intended learning outcomes, to highlight adequacies and inadequacies in learning. Therefore, self-referenced and criterion-referenced judgements are preferred to norm-referenced judgements as the emphasis is on improvement of achievement. But this approach is suitable only at the school (not board) level where the teachers through classroom testing aim at improving students' learning. As such the teachers can and should analyse classroom tests using this approach so that they can portray not only the students' achievement in terms of factors and conditions affecting their achievement but also identify the gaps and inadequacies in the teaching-learning process. Which of the two approaches may be used to form judgements depends on the purpose and focus of measurement.

6. How to Analyse?

Since the boards are concerned with public examinations only, it is safer to assume that the psychometric approach would be easier to follow. However, the first thing is to determine the quality of the measuring instrument itself. If the question paper does not itself reflect adequate reliability, further analysis of data that accrue from the use of such a tool is of no avail.

6.1. (a) *Finding Reliability of the Test*

There are a number of methods to calculate reliability of tests. In case of question papers having items or questions which can be marked on 1—0 scale as in the case of multiple choice or very short answer questions, one can easily use the split-half method or the Kuder Richardson formula (20) for the details of which the reader can refer to any book on educational statistics. For an easy-to-use formula, given by J.L. Saupe is quite appropriate and gives good approximation to K-R(20). Moreover, class teachers can also make use of this formula for quick calculation of the reliability of their tests. The formula is

given as under:

(i) *J.L. Saupe* (2)

$$r_{tt}=1-\frac{0.19\,K}{(S.D.)^2}$$

Where r_{tt} is reliability coefficient
$(S.D.)^2$ is the variance of scores
K = number of items on the test

(ii) *Mc Morris* (3)

Another formula is by Mc Morris (1972) which is fairly accurate.

$$\text{Index of measurement efficiency IME}=1-\frac{0.16\,K}{(S.D.)^2}$$

Where:

K is the number of items in the test, and
S.D.^{2} is the variance of Scores.

For quick calculation of standard deviation, teachers can make use of the following formula by Paul Diedriech: (4)

$$S.D.=\frac{\text{Sum of } \frac{1}{6} \text{ of highest score} - \text{sum of } \frac{1}{6} \text{ of lowest score}}{\frac{1}{2} \text{ of the total number of students}}$$

When the number is small use N−1 in place of N.

Test, retest reliability is not possible in a public examination as the same test cannot be given twice. In case of an essay test the calculation of reliability is a difficult problem. It can be planned through well structured questions ensuring uniformity of scores by means of a detailed marking scheme and instruction to the examiners. Attempt should be made to present uniform tasks for framing the questions. If the questions are well structured and balanced in terms of objectives or the abilities tested, there is not much difficulty in using the split half method of finding out the reliability of easy tests. In optional questions, especially when a free option is given, the difficulty is further enhanced—to split the test into two halves. However, if all questions are compulsory or balanced internal options are provided say, in a six questions essay test, we can obtain separate scores for each test i.e. for the first, third and fifth questions and also for the second, fourth and sixth. The correlation coefficient between the two halves can then be computed and followed by correlation using the Spearman's. Brown formula as illustrated below:

(iii) *Spearman's-Brown Correction Formula* (5)

$$r_{xx}=\frac{2r_{\frac{1}{2}\frac{1}{2}}}{1+r_{\frac{1}{2}\frac{1}{2}}}$$

Where r_{xx} = reliability of the total test
$r_{\frac{1}{2}\frac{1}{2}}$ = reliability you obtained by correlating half of the test with the other half (suppose 0.6)

Example

Suppose $r_{\frac{1}{2}\frac{1}{2}}=0.60$

Then using the above formula we have $r_{xx}=\frac{2(0.60)}{140.60}=\frac{1.20}{1.60}=0.75$

(iv) *Kuder-Richardson Formula-21* (6)

Using simplified Kuder-Richardson Formula-21 we have:

$$r_{xx}=1-\frac{\bar{X}\,(K-\bar{X})}{K\,(S.D.)^2}$$

Where $\bar{X}$ = mean
K = Number of *items* in the test
$(S.D.)^2$ = Variance (Standard deviation-squared)

Example

Suppose $\bar{X}=50$ S.D. $=6$ K $=60$ items

Then:

$$r_{xx}=1-\frac{50\,(60-50)}{60\ (6)^2}$$

$$r_{xx}=1-\frac{50\ \times\ 10}{60\ \times\ 36}\ \text{OR}\ r_{xx}=1-\frac{500}{2160}$$

Therefore, $r_{xx}=\frac{83}{108}=0.77$

The only way to compute the reliability of a test is to give two tests of the same ability and correlate the two sets of scores. This is true of essay tests also, in which questions receive different number of points. K.R. Formula-20 applicable to objective type questions is given below:

(v) *K-R Formula-20*

$$r_{xx}=\frac{n}{n-1}\left[1-\frac{pq}{\sigma^2}\right]$$

When:

n=number of items in the test
p=proportion of students who answer the items correctly
q=Proportion of students who answer the items wrongly=(1−P)
pq=variance of a single item scored as right/wrong.
σ^2=variance

(vi) *Chronbach co-efficient-alpha*

Essay type examinations—Co-efficient alpha developed by Chronbach (1951) is a generalisation of KR-20 formula (7) when items are not scored dischotomously i.e. right/wrong. For all compulsory questions type the following formula is usable:

$$r=\frac{n}{n-1}\left(1-\frac{\sum_{i=j}^{n}\sigma_j^2}{\sigma_0^2}\right)$$

Individual question variance may be found out and approximation S.D. calculation formula can be used.

Where is the variance of a single item. Where choice is provided Nuttal (1969) formula can be easily applied.

$$r=\frac{1}{I-1}\left[I-1\left(\sum_{j=i}^{k} n_j\, s_j^2\right) / s_x^2 \sum_{j=1}^{k} n_j\right]$$

Where: I=number of questions to be attempted
k=number of questions
n_j=number of students who answered question j
s_j^2=Variance of scores on question j
s_x^2=Variance of total scores.

6.2. *Item Analysis*

Simplified methods

The purpose of item analysis is to evaluate instructional effectiveness, estimate individual and group attainment through information about student's performance on each item and to improve test quality i.e. whether the items are doing what they are supposed to do. For this two important characteristics are, the finding out of the Facility Values (F.V.) and Discrimination Index (D.I.) of each item or the question. The following methods can be used:

(a) *Facility Value (F.V.)*

A. *Objective type items*

$$\text{(i) F.V. of an item} = \frac{\text{No. of students answering the item correctly} \times 100}{\text{Total number attempting the item}}$$

e.g. $\frac{40 \times 100}{50} = 80$ (F.V.)

(a) If number of students is less than 40 take the whole population.
(b) If number of students is between 40 and 100 top 27% and bottom 27% may be considered.
(c) If number of students is beyond 100 then top—bottom 10% will be sufficient.

(ii) Facility index can also be calculated by counting the number of students who attempt the item correctly from both, the upper and the lower groups i.e. 27% from each by using the following formula:

$$\text{Facility index} = \frac{\text{Number of correct answers in both the groups}}{\text{Total number of papers in both groups}}$$

	High Scoring (H)	Low Scoring (L)	Total
Correct	8	3	11
Incorrect	2	7	9
Total	10	10	20

$$\text{Facility index} = \frac{11}{20} = 0.55$$

B. *Essay-type (Multiple-point items)*

(i) *D.R. Whitney and D.L. Sabers.* We can calculate as under:
(a) Arrange papers from highest to lowest scoring students (Total Score).
(b) Highest 27% (or 25%) and lowest 27% (or 25%) are taken out for item analysis.
(c) Calculate sum of points made on each item for each of the two groups.
(d) Use the formula given below: (8)

$$\text{Index of difficulty} = \frac{\Sigma SH + \Sigma SL - (n_T)(X_{min})}{n_T (X_{max} - X_{min})}$$

Where:

ΣSH = Sum of the F X Column for high scoring group

ΣSL = Sum of the F X Column for low scoring group

X_{max} = Maximum possible score on the item

X_{min} = Minimum possible score on the item

n_T = Total number of papers in the combined high and low groups.

Example

TABLE—1

High Scoring group (H)			Low Scoring group (L)		
Item score	F	FX	Item score	F	FX
5	2	10	5	0	0
4	3	12	4	1	4
3	4	12	3	4	12
2	1	02	2	2	04
1	0	00	1	2	02
0	0	00	0	1	00
Total:	10	36	Total:	10	22

Using the formula we have :

$$\text{Facility Index} = \frac{36 + 22 - (20)(0)}{20(5-0)} = \frac{58}{100} = 0.58$$

The facility index is equal to 0.58 which means that the class average is 58% of the difference between the lowest and the highest possible class average.

(ii) *Morrison index*

$$\text{F.V. of a question} = (50 + M_Q - M_T)$$

Where: M_Q = Mean percentages mark on that question for all candidates

M_T = Mean ability in percentage total mark of those who attempted the question.

(a) use entire population if up to 40.
(b) use top-bottom 27% if 40 to 100.
(c) use top bottom 10% if population is > 100.

(iii) *Willmott Nuttal*

$$F.V. = M + MQ - MT$$

Where:

MQ = Same as above
MT = Same as above
M = Mean % marks.

(iv) *Edwin Harper*

$$F.V. = \frac{MU + ML}{2 \times \text{Max. marks for the question}} \times 100$$

Where:
MU = Mean mark of top 1/5th of students on the question
ML = Mean mark of bottom 1/5th of students on the question

(b) *Discrimination index*

Index of discrimination is an important indicator of item effectiveness for any discriminating test. An item which is attempted correctly by students who have good command of the objectives measured by the total test and which are failed by students who do poorly on the total test, is considered to contribute in a positive way to the ability of the test to discriminate accurately among different levels of students performance, vis a-vis the relationship of the test scores. The index of discrimination can be calculated by dividing test papers into high scoring and low scoring groups.
Following are some of the formulae in use:

(i) Index of discrimination $= \frac{RH - RL}{NH}$

Where: RH = number of right answers in high scoring group.
RL = number of right answers in low scoring group.
NH = number of papers in either high (H) or low (L) Scoring group.

(ii) *Johnson upper-lower index*

D.I. = (F.V.) Higher ability 27% (F.V.) lower ability 27%. This

is quite useful for classroom tests.
(a) if strength is < 40 grouping of 50%—50%.
(b) if strength is between 40-100 grouping of 27—27%.
(c if strength is > 100 grouping of 10%—10% can be used.

(iii) *For choice type* (Essay)
Harper's Formula

$$D.I. = \frac{1.8\ (M_U - M_L)}{\text{Max. marks for the question}}$$

Where: MU = mean mark of upper 1/5th scores
ML = mean mark of lower 1/5th scores

(iv) D.I. is the product moment correlation between marks on the question and the total marks.

$$D.I. = \frac{\Sigma\ (x-\bar{x})\ (y-\bar{y})}{N\ S_x\ S_y}$$

Where: x = mark on question (whose D.I. is to be found) of a given student
y = total marks of the same student
N = number of students
S_x = sd. of x
S_y = sd. of y

(v) *Whitney and Sabers formula*

$$\text{Index of discrimination} = \frac{S_H - S_L}{n\ H\ (X_{max} - X_{min})}$$

Where:
S_H = sum of the FX column for the high scoring group
S_L = sum of the FX column for the low scoring group
n_H = number of papers in the high scoring group (or low Scoring group)
X_{max} = maximum possible score on the item
X_{min} = minimum possible score on the item

In the previous example (Table-I) $S_H = 36$, $S_L = 22$, $X_{min} = 0$ $n_H = 10$

Therefore, $D.I. = \frac{36-22}{10\ (5-0)} = \frac{14}{50} = 0.28$

Depending upon the purpose and nature of data that accrue from a testing situation one may use the relevant formula to

compute facility indices and discrimination indices. Individual items may be very good when framed but when grouped together in the form of a test they may behave differently. An item having very good (50%) facility index may come out as poor discriminator with a low discrimination index. If a board undertakes this sort of analysis to find out the quality and behaviour of each question then it is in a better position to take necessary steps for the next year to:

(a) improve the quality of questions.
(b) improve the validity and reliability of the testing instrument.
(c) feedback to the item writer/paper setters to improve their skills.
(d) feedback to the schools to improve their classroom tests.

Besides the question and the question paper, equally important is the need to analyse responses of students. What are the types of mistakes students make? What are their strengths? What type questions do they attempt more satisfactorily? What are the instructional objectives which demand more appropriate answers? What are the content areas in which students performance is not satisfactory? On which form of questions is students' score more compared to other forms? What are the questions which created trouble for examinees to know the nature and scope of expected responses? These and many more questions can be identified, the answers to which can be found out by analysing students' responses. Besides analysis of students' responses, the test results should also be analysed. Frequency distribution of scores, mean, median and mode, dispersion etc. are the basic requirements which can be easily applied to all variety of tests, theory and practical, so that the scores could be transformed and made comparable. How rigorous the techniques they are able to use is upto the Boards. At least simple transformation of raw scores into Z-scores is quite easy to apply and on the basis of which we can convert marks into T-score also. What is the underlying emphasis to ensure comparability of marks in theory vs practical, external vs internal, written vs oral, History vs

mathematics etc. Involves only an administrative decision and a statistical treatment of data.

7. Using Results of Analysis

The simple analysis of student's performance which can be utilised with profit pertains to the analysis of test results in terms of:

(a) instructional objectives.
(b) content areas.
(c) students' errors.
(d) distractors' plausible (multiple choice items)

Once analysis is done carefully the data can be used for following purposes:

1. To judge the efficacy of the various instructional objectives.
2. To discover the abilities (intended learning outcomes) which are not properly developed.
3. To identify the content areas which may be redefined or omitted from the curriculum.
4. To take decisions about the grade placement of certain concepts.
5. To suggest possible teaching-learning strategies for improving instructional practices.
6. To provide guidelines to the institutions for improving evaluation practices.
7. To undertake further diagnostic studies for detailed analysis of errors.

9. Use of Item Analysis

The discrimination index indicates defective and ineffective items, thereby providing useful feedback to the item writer to improve his items. Over the years a teacher can develop a pool of individual items which can be typed on 10cm × 15cm cards along with the item indices. The labour of preparing good items decreases after a few term tests as the main job of the teacher

would then be of selecting, revising and assembling of items rather than preparing new items for the test. The item selector is often tempted to discard defective items, although improvement is possible in most of these items if care is taken. A study by Lehmann and Mehrens indicated that revising of a item takes only 1/5th of the time needed to write a new item. It is therefore, necessary that more effort be made to improve the existing items and when replacement is to be made, the concept and objective tested by the item urgently should remain the same while making the replacement of items.

The characteristics of the items change drastically after the curriculum changes. The facility indices and discrimination indices change from class to class. Therefore, a teacher should be cautious in interpreting the data from a single item as indicator of an individual student's achievement.

Paper setting is only meant for collecting information about students' learning. Improving the validity and reliability of the evaluation instrument is only an attempt at getting relevant, adequate and dependable evidence. How best this evidence can be used to improve students' learning, instructional strategies, measuring instruments and management of examinations is equally important, for this concept of evaluation as feedbaek must be appreciated. Once the ecology of testing is realised by the boards in particular and teachers in general, more and more emphasis on the feedback aspect of evaluation would be laid to bring about improvement in instructional objectives content or education, methodology of teaching, technique of testing and even management of examination. This would, however, not be possible unless the teachers in general and managers of examinations in particular take cognizance of the role of analysis and feedback in the total evaluation system.

CHAPTER XVII

EVALUATION TRENDS IN 10+2 PATTERN OF SCHOOLING

1. Introduction

The curriculum for the ten-year school prepared by the N.C.E.R.T. (1975) has provided the basic framework for improving the quality of education at different stages of school education. It is for the first time that guidelines for a broad based curriculum have been suggested keeping in view the societal needs for general education. This framework, popularly known as the N.C.E.R.T. approach-paper on curriculum, (1) has been widely discussed at the national and state levels. By and large educationists, curriculum specialists and teachers have accepted this pattern of schooling as the most relevant framework for educational reconstruction.

In this chapter attempt is made to highlight trends of evaluation in the context of 10+2 pattern of schooling. The reader after going through this chapter should be able to

(a) identify the assumptions underlying the ten-year school curriculum.
(b) formulate key concepts forming the framework of the new curriculum.
(c) trace various implications of evaluation in the context of curriculum intentions.
(d) make generalisations about the changed emphases in the evaluation process and the system as envisaged in the curriculum.

(e) analyse the philosophical, social, psychological, scientific and economic bases of trends in educational evaluation.
(f) analyse the conditions and developments in the field of education that have bearing on trends in evaluation.
(g) infer various trends relating to different aspects of evaluation.

Since this section is aimed at developing a framework for evaluation it is not necessary to discuss the highlights of the new curriculum. However, the framework of the curriculum includes evaluation as its integral part. As such, the evaluation system is embedded in the curriculum roots of the new pattern of schooling. Assumptions underlying the new curriculum provide the basis for identifying the assumptions of the framework of evaluation. It is, therefore, necessary to state these assumptions in clear-cut terms.

2. Assumptions

Though the approach-paper does not list as such the underlying assumptions, it is obvious from the content of the document that the disciplines of the curriculum, the objectives, the methodology of teaching and the mode of evaluation as reflected in this document, appear to be based on the following assumptions: (2)

(a) A good curriculum is one which fosters a national identity and reflects national development.
(b) Uniformity of the curriculum would lead to reduce the gap between standards of education among the different states.
(c) Curriculum renewal at regular intervals is essential to meet the demands of new knowledge and the needs of the society.
(d) Teaching of mathematics and science is essential for every citizen to develop a rational outlook.
(e) Teaching of work-experience is essential to reduce the gap between the elites and the masses by relating education to the world of work.

(f) Language is a powerful tool for developing a catholicity of outlook and a means to cultivate the basic value of culture.
(g) Artistic experience, health and physical education and character-building are essential components for self-actualization of an individual.
(h) Flexibility in curriculum provides for better individual development and motivation.

3. The Key Concepts

It is for the first time that Science, Mathematics, Work-Experience, Health and Physical Education have been introduced as compulsory subjects as a part of general education for the ten-year schooling. Multiple-entry, the semester system, the unit approach, advanced-level units, evaluation and feedback etc., are some of the new concepts which are stressed in the new curriculum. It will not be out of place to suggest here the four key-concepts which form the warp and woof of the new curriculum. These may be termed as compositeness, integration, learnability and feedback which are further explained below.

3.1. *Compositeness*

It refers to the total picture of the disciplines of the curriculum at various stages of education and the different components of curriculum. Emphasis in the new curriculum would, therefore, not be limited to the content of the curriculum but also of all the four major curriculum components, namely; objectives of the curriculum, content of the curriculum, the methodology of the curriculum and evaluation of the curriculum. Development of the curriculum of a particular discipline is, therefore, envisaged in accordance with the total curriculum plan.

3.2. *Adaptability*

This term refers to the flexibility in curriculum planning, development, implementation, and evaluation of curriculum. It is visualised that each state will have the freedom to formulate

the curriculum objectives, curriculum content, curriculum strategy and curriculum evaluation in accordance with their educational antecedents, expected inputs and the needs of the state. It calls for emphasis on the individual, learner-based methodology of teaching and mode of assessment. It highlights the need for individualised instruction and individual-based evaluation. The unit-teaching approach would be the basis for instructional inputs which would emphasise self-based learning and self-referenced evaluation.

3.3. *Integration*

Integration is viewed in terms of articulating the curricula of different stages on school education as well as the various disciplines of the curriculum. It also refers to integration of teaching and testing for adopting the unit approach of teaching and testing. Likewise, organisation of learning experiences and performance measures would be objective-based. The emphasis would be on the need for appreciating the dynamic interaction between objectives, content, teaching-learning strategies and evaluation.

3.4. *Learnability*

This aspect emphasises the need for cultivating the habit of self-learning and self-evaluation, development of intellectual skills, cultivation of a rational outlook. Emphasise on the concept-attainment principles and generalization would find a more prominent place in the educational process. Factual information and knowledge of specifics would no longer be emphasised.

3.5. *Feedback*

This aspect refers to providing pupils the information about adequacies or inadequacies in their learning. Diagnosis of pupils' difficulties, application of correctives, identification of instructional deficiencies and efficacy of concepts and objectives would attract more attention on the part of teachers and evaluators.

From the above-mentioned key concepts, one can easily infer that the emphasis in the new curriculum is on understanding the closely inter-related curriculum components, thereby emphasising the need for making evaluation integral to the total process of curriculum development.

4. Implications for Evaluation

With reference to the first concept, it can be assumed that when evaluation is considered an integral part of the teaching-learning process it brings quality-control into the whole process of education. Therefore, in the new curriculum, the emphasis would shift from pupil's evaluation to curriculum evaluation. Likewise evaluation would be extended from the cognitive to all other non-scholastic areas.

When evaluation is done in terms of learners, rather than the class, it provides more valid measures of students' learning and a more reliable basis for adapting instruction to the needs of the learner. Accordingly, the emphasis would shift from achievement test to diagnostic tests, from mixed-ability teaching to individualized instruction and from norm-reference testing to criterion-referenced testing.

When evaluation is based on the objectives of the instruction and the learning experiences provided in a unit, it provides evidence about pupil's learning that is likely to be more reliable. This is based on the assumption that the teacher who teaches is the best judge of his students. This would lead to the use of a wider variety of tools and techniques, cooperative evaluation, objective-based and integrative teaching and testing.

In terms of feedback, it is assumed that regular feedback about pupil's achievement would improve their achievement. Accordingly, frequent evaluation at regular intervals, unitwise teaching and testing, diagnosis of pupils' weaknesses, remedial instruction, proper interpretation and reporting of results immediately after evaluation will be the more sought-for activities of the teachers.

When evaluation is considered an integral part of the teaching-learning process, evaluation must be relevant to the objectives of teaching as well as to the learning experience provided to

achieve those objectives. Evaluation therefore, is based on these two aspects of the educational process and at the same time provides evidence about the efficacy of the two aspects. Intended learning outcomes, collection of evidence, reporting of evidence and the use of evidence can be visualized as the five major aspects of evaluation.

When we think of the intended learning outcomes of a particular unit of teaching, we have to specify the objectives of that unit and even decide about the level of learning in that particular objective which the student is expected to achieve after instruction. Besides, how many students do we expect to attain that level? Apart from this, the time-limit for obtaining the intended level of achievement can also vary from learner to learner. Therefore, in the new pattern, there is flexibility in the expected level of learning in terms of the time required to achieve that level. Students can take their own time to reach the expected level of achievement.

As for collection of evidence about pupils' growth, the significance of the data-gathering devices cannot be overemphasised. Similarly, the reliability of the data-gathering devices, the usability of the evidence and its systematic recording are the main features to be reckoned with. It may be pointed out here that the relevance of the evidence, rather than the reliability of the evidence, is more important in teacher-made tests.

For analysis and interpretation of evidence, evaluation should be more concerned with the objective-wise performance. Norm-referenced measures cannot, therefore, be considered as important as criterion-based measures in the examinations of tomorrow.

For reporting of evidence, what is intended is to indicate the level of performance rather than a judgement on the performance. Performance cannot be taken as an absolute measure but only as a relative performance which is, in fact, only an estimation and not as a perfect measure. Thus the focus of evaluation would be on diagnostic appraisal rather than on labelling of a student; on improvement of his achievement rather than on measurement of it. Judgement is revisionary in nature rather than irreversible, as we find in today's public examinations.

pupils' growth in different aspects of their development is reported to students, teachers, parents, employers and institutes of higher learning, the more appropriate and reliable would be the decision taken for classification, certification and selection of students for different purposes. Therefore, regular recording of pupils' performance in various areas of development is a pre-requisite for every evaluation programme.

6. The Bases of Trends

This section deals with the trends in evaluation. That change in emphasis on evaluation is conditioned by the shift in educational theory, is not difficult to conceive. Moreover, it is also not difficult to establish a relationship between a trend and the social need. Like a fashion which is not only the outcome of a social need but also a craze for deviation from the normal, a trend does not necessarily reflect the fulfilment of social needs. It may receive social support or engender hostility and criticism. Nevertheless, a trend does depend on the cumulative pressure built against the unwanted tradition which is alleged to have stopped meeting the educational need of the time. The same is applicable to evaluation as well. As long as evaluation is considered as the intelligence service of the teaching-learning process, its emphasis would continue to vary with the educational process but the longevity of the trend would depend on the utility and diffusibility of the change which indeed stems from various predisposing factors.

6.1. *Philosophical Basis*

A system of evaluation, vis-a-vis the educational process is a reflection of the philosophical, societal, psychological scientific and ecological bases of the educational system. Therefore to trace the trends in evaluation we have to identify those factors which necessitate the change under each of the above mentioned bases of education. The democratisation of the process of education has led to the recognition of the worth of an individual and his role in

the decision-making process. His ability to participate in the learning process with his assets and liabilities cannot now be questioned. The purpose of evaluation has to be thought of in terms of the values of the system of which it is an integral part. Evaluation as a service component rather than as a judgemental process has, therefore, to be appreciated. Thus evaluation must take into account such skills and abilities which are conductive to the development of the individual.

6.2. *Social Basis*

From the societal point of view, of late there has been unscathing criticism of the utilisation of public funds on educational investment. To what extent expenditure on big curriculum projects, like B.S.C.S. has been commensurate with the dividends that accrued from it, was a question raised by the public in U.S.A. Is the content of education, vis-a-vis inclusion of various subjects, in accordance with the social needs? Can we really pass irreversible judgements without having a negative impact on examinee's personality? Are examinations really economical? All these and such other questions have created among administrators the need for accountability of all new educational endeavours. So the evaluation machinery cannot remain oblivious of the need for cost-benefit analysis in any programme of educational evaluation.

6.3. *Psychological Basis*

From psychological angle, there is increasing realisation of the role of objectiveness in teaching and its impact on learning. More and more students' involvement, individualised, self-paced learning, diagnostic teaching and learnability are, among others, the preferred modes of instruction and effective learning. An evaluator of today and tomorrow would, therefore, be compelled to gear his evaluation to the appraisal of such abilities using appropriate techniques and relevant situations.

6.4. *Political Basis*

Politically, education in developing countries is no longer the monopoly of the elites but a birth right of each citizen irrespective of cast, creed and class. The needs of weaker section of society are considered on priority basis. Differentiated curricula, graded textbooks and diagnostic evaluation are receiving more and most attention in democratic countries. Agencies conducting external examinations are facing problems of de-emphasising these examinations and delinking them with employment, giving way to teacher-based continuous assessments.

6.5. *Scientific Basis*

Scientific basis of measurement are now questioned more for their reliability and relevance to the goals of education. Grading rather than numerical scores is gaining ground. The need for well designed question papers, scoring objectivity, comparability of results and sound judgements are the demands raised as a sequel to the onslaughts of misplacement and misclassification of students resulting from the present tools and techniques of evaluation.

6.6. *Ecological Basis*

Lastly, the impact of examination on the various components of the teaching-learning process especially the pathology of examinations is now attracting more attention of the evaluators and educationists than before. The use of evidence resulting from evaluation procedures, for adapting ends and means to each other, is now being considered more or less obligatory, to improve students' achievement and instructional practices. To ward off the strangulating influences of examinations on the teaching-learning process there is need for understanding the interaction of evaluation with other components of the teaching-learning process. Thus the present day evaluators are becoming more and more cognizant of the ecology of examinations.

7. The Underlying Principles

From the above-mentioned bases viz. the philosophical, the societal, the psychological, the political, the scientific and the ecological, one may be tempted to identify the underlying principles which would form the basis for tracing the various implications and trends in evaluation. Such trends which emerge from these principles may relate to various phrases of evaluation such as collection of data, analysis, interpretation, judging and decision making besides the very concept of 'evaluation' itself. The following principles may be delineated: (3)

(a) A knowledge of inadequacies in learning motivates the learner to feel the need for improvement in learning.
(b) A knowledge of expected learning outcomes provides direction for teaching and learning.
(c) A continual diagnosis of students' gaps in learning accelerates students' learning if proper remedial measures are taken regularly.
(d) The more comprehensive the evidence about pupils' growth the more sound and the judgements made about learners and introduction.
(e) If the learner is judged in terms of his own assets and liabilities and his adequacies in learning are continuously reinforced, he develops a positive self-concept which improve his learning.
(f) Regular feedback of test-results generates a dynamic interaction among the different components of the educational process which leads to homeostasis in the learning milieu.
(g) Good evaluation practices promote better learning and improve teaching.

The above mentioned assumptions pave the way for identifying the trends in educational evaluation, some of which are tangible while others are intangible.

8. The Emerging Emphases

(a) *The Broadened Scope*

Recent literature in the field reveals that evaluation is no longer limited to collection of data. Its scope is being extended to description of measurement for forming judgements on the basis of which decisions are made about the learner and the learning process. Measurement is a pre-requisite to any evaluation but not synonymous with it. The focus of evaluation is on improvement of students' achievement through timely provision of data about their inadequacies in learning.

Therefore:

The restrictive concept of measurement is broadened to that of evaluation thereby focussing more attention on the decision-making role of evaluation (Pedagogical) than merely on its measurement role. (4)

(b) *Objective Centredness*

Since Bloom's taxonomy (1956) followed by Krathwhol's (1964) and Harrow's (1972) emphasis on objective based teaching and testing is on the increase. Despite some difficulties, the objective-based approach to evaluation has come to stay. Specification of objectives in terms of behaviour is now considered useful in providing direction to teachers for both teaching and testing. Both norm-referenced and criterion-referenced models of teaching emphasise the need for unit objectives against which teaching and testing are validated. The content is not the objective of testing but a vehicle or medium of testing various skills and abilities.

Therefore:

Emphasis on testing of content is shifted to testing of objectives of discipline thereby highlighting the need for appraisal of intended learning outcomes of a learning unit through the medium of prescribed content.

(c) *Formative Function*

Evaluation is not a one shot affair at the end of a course, a term or a semester. It is considered now a service component of the teaching-learning process involving a continuous appraisal of pupils' growth with a view to adapt instruction to improve students' learning.

Therefore:

Emphasis from summative evaluation has shifted to the formative and diagnostic evaluation in order to make evaluation a quality control device and a service component

(d) *Total Development Directed*

Recognising the integrative development of the child in the area of cognitive, affective and psychomotor domains the present evaluation procedures limited to appraising of cognitive development cannot be considered adequate to appraise the total child. Since affective entry behaviour enhances learning and psychomotor skills act as a limiting factor besides providing motivation and interest, their role in the process of learning cannot be overemphasised.

Therefore:

Emphasis in evaluation has extended from mere appraisal of cognitive development of the child to the whole child to encompass the total growth of pupils, in both scholastic and non-scholastic areas.

(e) *Illuminative or Descriptive Evaluation*

Because of the comprehensive scope of evaluation to appraise the total child, it becomes imperative to evaluate pupils in all such situations which are relevant to the gathering of evidences relevant to the area of scholastic and non-scholastic traits. Thus the learners' peers, parents, community, teachers and the learner himself are the relevant sources of evidences which have to be tapped in the context of the social milieu which conditions learning. The mere psychometric data collected from

the experimental approach will, therefore, not do. Some sort of a social-anthropological approach is called for which could illuminate the total array of agents, relations and contexts so as to reflect the impact of the total learning milieu on students' learning.

Therefore:

> The traditional Agricultural-Botany paradigm involving experimental design of evaluation is giving way more and more to the descriptive social anthropological model that takes into account the context, the social set up and participants' observations which have a direct bearing on illuminating the learners' achievement and growth. (5)

(f) *Diagnostic Focus*

The shift from prediction to diagnosis is evident now, especially in the context of the recognition of weaker section of students. Research-based encouraging results through remedial programmes undertaken as a sequel to diagnostic testing have resulted in giving more emphasis on diagnosing weaknesses than on measuring achievement for grading. Diagnosis, remediation and improvement of achievement is the focus.

Therefore:

> The stress of todays' measurement is reflected more and more on identification of students' difficulties and errors rather than on measurement of achievement in a given content. The focus of evaluation is now on improvement of students' achievement rather than on measurement of achievement.

(g) *Relevance of Evidences*

Once we accept the role of illuminative evaluation in the educational scene we are bound to gather evidence about various aspects of students' development through different sources and personnel who watch the learner grow through the corridors of time. This would require not only the use of a wide variety

of tools and techniques which could portray the evidence of each type but also the opinion of all those who are concerned with the growth and development of the pupils. Techniques like those of observation, rating scale, checklist, sociometric devices, performance tests etc. would now find a more prominent place in educational evaluation. This would also generate evidences that are more relevant to different aspects. Process of performance will in no way be less significant than the end products of learning.

Therefore:

The relevance of measurement is being considered more important than the reliability of measurement as far as the class tests are concerned. More and more emphasis is given to the judgement of persons rather than tools and techniques they use.

(h) *Processing for Meaningful Scores*

The methodology of processing is at present limited to arbitrary cut scores and divisions into which students are classified on the basis of raw scores. This has caused difficulty in comparing students' performance from subject to subject and from examiner to examiner. The need for scientific interpretation in combining the scores is increasingly felt now in order to pass sound judgements. Likewise, the need for use of grades in place of marks is considered more reliable measure for minimising misclassification of students when viewed in the light of low reliability of existing instruments of measurement.

Therefore:

The methodology of processing the evidence rests more and more on subject-wise analysis of results which are more diagnostic than global scores. Transformation of raw scores into standard scores for combining marks and their scaling is preferred to arithmetic totalling for enhancing the meaningfulness of scores. Alternatively, use of grades in place of marks are recommended more strongly both in internal and external examinations.

(i) *Criterion-referenced Judgements*

At present group-based (class) or norm-referenced judgements are common. At public-examination level, the irreversibility of judgements as pass or fail or third division, have an adverse effect on students' learning and development. Declaring students as deviants, failures and under-achievers with reference to a norm group has a corroding effect on the personality of the students. Every student benefits in accordance with his aptitude, time allowed to learn, perserverance, quality of instruction and his ability to benefit from instruction. As such evaluation must look for reference to the child himself as far as progress is concerned. Likewise the achievement on the predetermined criteria of success on a topic or a skill is more important than knowledge of his deviation from the norm group. (5)

Therefore:

Self-referenced and criterion-referenced measurement rather than norm-referenced measurements; mastery judgement rather than deviation judgements; reversibility rather than irreversibility of judgements, are the significant departures sought for by the modern evaluators to provide sound basis for academic and administrative decisions.

(j) *Feedback for Equilibrium in Teaching-Learning Process*

Feedback and correctives being essential features of any evaluation programme, the examining agencies and class teachers are now more conscious of the need for feedback of results to the institutions and students respectively. In fact very little work is available on the ecology of examinations. But there is a growing consciousness among the teacher-evaluators and academic administrators to use the cibernetic approach of system analysis whereby the impact of one component of evaluation on other components in a given system or educational process could be known by understanding the interaction among each other. How can we use evidence for generating a dynamic interaction to energise the teaching-learning system by using evaluation as feedback, is the emerging interest of evaluation experts of today.

Therefore:

> The pathology of evaluation practices spread through the use of evaluation as a judgemental device that leads to undesirable influences on teaching-learning process is being sacrificed now at the altar of the ecology of evaluation which aims at dynamic equilibrium by adjusting the teaching-learning strategies to improve students' learning (7).

In the second part of this chapter attempt has been made to highlight the significance of the philosophical, psychological, societal, political, scientific and ecological basis of educational evaluation. Each of these bases pinpoints some nodal issues which become the basis of change in the emphasis on evaluation practices. However, all such issues or foci are based on certain assumptions relating to the learner and the learning process. Though some people may question these assumptions, yet these do reflect some consensus. From these assumptions, follow the various trends or emphases which we find today reflecting the needs, values, preferences and promises. To what extent these trends are acceptable, relevant and practicable depends on the faith that our clientele develops, the availability of resources and the potential energy of the innovation. For a trend to be accepted, stabilised and reflected in the system we have to do valiantly, hope confidently and wait patiently, the results of our endeavours, like a good researcher.

10. Need of the Day

Once the role of evaluation in teaching and learning is appreciated by the teachers, students and administrators, the importance of external examination would be relegated a secondary position. Such examinations can be used to compare the standards of different institutions, to select students for admissions and jobs. When schools become the agency for conducting examination and teachers the agent of this change, the validity of students' evaluation cannot be challenged. Then teachers who teach would assess the performance of their wards. Students would not devise ways and means to get grades.

bringing quality control in the teaching-learning process.

(f) develop better insight into the dynamic role of educational evaluation for improving pupils' learning as well as the teaching-learning strategies.

(g) identify the reason and issues related to delinking of degrees in certain jobs.

(h) appreciate the need for judging pupils' performance, curriculum effectiveness, teachers' accountability and programme evaluation in the context of total school of evaluation vis-a-vis the learning milieu and social setting.

(i) understand the role of National Testing service in comparing and setting standards.

2. Evaluation Process and Examination Reform

Under article 8.23 of the N.E.P. "assessment of performance is regarded as an integral part of any process of learning and teaching. Apart of sound educational strategy, examinations should be employed to bring about qualitative improvements in education."

This recommendation envisages the importance of examination reform as a means to improve teaching-learning strategies as well as the improvement of the performance standards by bringing quality control in the whole process of education. This requires relating examination reform to:

(a) curriculum development process.

(b) teaching-learning strategies.

(c) development of instructional and learning material.

(d) mechanics of developing instruments of evaluation, conduct of examinations etc.

(e) learning milieu of the institution.

From the above mentioned relationships we may observe that a close and inter-dependent relationship exists among pupil evaluation, curriculum evaluation and institutional evaluation, focus of each being on improvement of students' learning, teaching-learning strategies and learning environments.

Article 8.24 states, that, "Objective will be to recast the

examination system so as to ensure a method of assessment is valid and reliable measure of student development that and a powerful instrument for improving teaching and learning. In functional terms, this would mean according to N.E.P.:

(a) the elimination of excessive element of chance and subjectivity.
(b) the de-emphasis of memorisation.
(c) continuous and comprehensive evaluation that incorporates both scholastic and non-scholastic aspects of education, spread over the total span of instructional time.
(d) effective use of evaluation process by teachers, students and parents.
(e) improvement in the conduct of examinations.
(f) the introduction of concomitant changes in instructional materials and methodology.
(g) introduction of the semester system from the secondary stage in a phased manner; and
(h) the use of grades in place of marks.

Under this article (8.21) two directions are indicated:

(a) To improve the quality of measuring instruments for more dependable evidences. In this regard focus is on:

    (i) maximising the objectivity in assessment.
    (ii) de-emphasising the role of memorisation in construction of instruments of evaluation.
    (iii) use of grades in place of marks.

(b) To improve examination system in such a manner that they become useful devices for bringing improvements in the teaching-learning system. For this focus is on:

    (a) phased introduction of semester system at secondary stage, thereby, ensuring regular study; better feedback, more intensive coverage of syllabus for examination and apportioning curriculum load in four semesters.
    (b) incorporating non-scholastic areas of pupils' development for more comprehensive assessment.

(c) continuous appraisal, diagnosis, remediation and feedback to in order to improve students' achievement and adjusting or adapting teaching learning strategies.
(d) diagnostic and formative function rather than on judgemental function of assessment, thereby, providing feedback to teachers, students and parents about pupils' adequacies and inadequacies in learning.
(e) developing mechanics of examinations scientifically to make examination marks, vis-a-vis results, more simple, meaningful and comparable.
(f) dove-tailing teaching-learning practices, instructional and learning materials, curriculum transactions and other collateral changes in the system by using evaluation as quality control device and service component of the educational process.

According to article 8.25 "the above goals are relevant both for external examination and evaluation within educational institutions. Evaluation at the institutional level will be streamlined and predominance of external examination reduced".

This article focuses our attention on:

(a) close relationship between external examination and school evaluation.
(b) developing a sound system of continuous comprehensive evaluation in schools to make teachers' assessments more valid and reliable.
(c) the need for making teachers responsible and accountable for developing cognitive as well as non-cognitive faculties of pupils' growth.
(d) reducing the dominating influence of external examinations on class room practices and in certifying pupils' achievement for promotions, failures etc.

3. Delinking Degrees from Jobs

Articles 5.38 to 5.41 point that "beginning will be made in delinking degrees from jobs in selected areas. The proposal cannot be applied to occupation-specific courses like Engineering, Medicine, Law, Teaching etc. Services of specialists with

academic qualifications in humanities, Social Sciences, Science etc. will continue to be required in various job positions".

Delinking will be applied in services for which a University degree need not be a necessary qualification. Its implementation will lead to a refashioning of job specific courses and afford greater justice to those candidates, who despite being equipped for a given job, are unable to get it because of an unnecessary preference for graduate candidates".

With respect to delinking of degrees the focus is on the need for:

(a) considering the candidates having needed skills or job training as qualified for appointment as apprentices for certain jobs where university degree is not needed for selection.
(b) re-designing all such courses which are job specific to make such candidates eligible without requiring a university degree.
(c) continuing the degree requirements in case of occupation specific courses like medicine, engineering, law, teaching etc.
(d) devising quality control mechanism in the form of nation-wide testing programme on voluntary basis.

4. Programme of Actions (P.O.A.)

Corresponding to N.E.P. the following measures are suggested in the P.O.A. in Chapter XVIII relating to evaluation process and examination reform.

4.1. *The Policy and Strategies for Implementation*

The policy visualises integration of the assessment of performance with the process of learning and teaching, and utilising the process of evaluation to bring about qualitative improvement in education (para 8.23). In order to ensure that the method of assessment of students' performance is valid and reliable, the following short-term measures are proposed:

(a) *At the School Level*:

(i) Public examinations will continue to be held only at the levels of Classes X and XII.

(ii) Decentralisation of the operation involved in the conduct of examinations to make the system work more effective.

(iii) School Boards in certain States have set up a number of sub-centres to decentralise the conduct of examinations. Adoption of similar measures by other State will be pursued.

(iv) In the event of decentralisation as indicated above, the State Boards of School Education would continue to get the question papers set and printed, consolidate the results of examinations and also undertake test checks on random basis of the functioning of the sub-centres; and

(v) Spot evaluation of answer scripts.

(b) *At the University Level*:

(i) Continuous institutional evaluation will be introduced at the post-graduate level, to begin with, in Unitary Universities, Demand Universities and Autonomous Colleges;

(ii) Students' performance will be indicated through letter grades, and assessment of overall performance will be on the basis of cumulative grade point average;

(iii) Provision will be made for improvement performances through subsequent appearances without involving any disadvantage to the candidate;

(iv) External examinations will continue to be held by universities which have a large number of affiliated colleges and efforts will be made to improve the conduct of examinations through effective decentralisation as indicated for school level examinations;

(v) Modifications in the qualifying recruitments for admissions in the universities and colleges will be examined to accelerate the process of change in the school level examinations.

(c) *Conduct of Examinations*:

(i) The possibility of introducing legislation to define various malpractices connected with examinations and to treat them as cognizable and unbailable offences will be considered;

(ii) Such laws will also, when enacted, make provision to prescribe the nature and type of punishments for various offences under the law, and to include within its scope persons engaged in various operations connected with examinations and also to provide protection of them; and

(iii) Innovations and experiments in the conduct of examinations, like printing and distribution of question papers with questions arranged in different sequences to avoid copying and other unfair means in the examination halls.

4.2. *Integrating Evaluation with Teaching and Learning*

In order to attain the objective of integrating the process of evaluation with teaching and learning, several long-term reforms will be necessary. For this purpose, the following programmes would be considered:

(a) *At the School Level*:

(i) The Boards of Education will lay down the levels of attainment expected at classes V, VIII, X and XII;

(ii) The Boards will also prescribe the learning objectives corresponding to these levels of attainment in terms of knowledge and comprehension, communication, skills in the application of knowledge, and the ability to learn;

(iii) Schemes of evaluation consisting of examinations to test those aspects of learning which can be assessed through formal examinations, and the procedure for assessing those aspects which cannot be tested through such an examination, will be developed. Abilities and proficiencies which can and should be assessed through institutional evaluation will be identified and procedures evolved for such evaluation;

(iv) The development of schemes of evaluation is a continuing process. To provide professional support to this process, the Boards of Education will consider setting up a Consortium for initiating research and development in evaluation procedures and in the conduct of examinations;

(v) For performing this task, the Consortium will adopt selected schools as pilot centres and will hold examinations and award certificates for the students of such schools;

(vi) Before question papers are set, a detailed design will be evolved indicating the weightage to be given to various areas of content, types of questions and the objectives of teaching/learning;

(vii) Along with external examinations, continuous institutional evaluation of scholastic and non-scholastic aspects of education will be introduced;

(viii) Evaluation of students' performance will move towards cumulative grading system;

(ix) In the big States, the possibility of establishing more than one Board of Education will be considered, so that the number of students to be examined by one Board does not exceed one lakh; and

(x) Procedures will be developed for the appointment of Chairmen/Secretaries of Boards of Education and Controllers of Examinations to inspire confidence among public.

(b) *At the University Level*:

(i) The possibility of developing alternate system of evaluation in place of external examinations for affiliated colleges will be explored;

(ii) The question of some universities functioning only as examining bodies for a number of colleges will be examined;

(iii) Academic reforms visualised in the policy like flexibility in the combination of courses, modular structure, provision for accumulation of credits, redesigning of courses, etc. will lead to considerable decentralisation

in the evaluation process. Detailed schemes will be evolved to facilitate transition to new evaluation procedures concurrently with the changes in the content and structure; and

(iv) An agency w ll be developed either as part of the AIU, or independently, for continuous research and development in evaluation procedure.

4.3. *General Recommendations*

(a) Integrity of the examiner is crucial to the credibility of the examination system.

This credibility can be established by the openness of the examinations. It has to be recognised that students have the inalienable right to scrutinise their answer scripts and its evaluation and also compare them with those of others;

(b) The practice of declaring results in terms of over-all divisions and pass/fail may be reviewed and substituted by a system of declaration of results in terms of marks/grades in each subject separately;

(c) Candidates should have the opportunity to improve upon their grades through subsequent attempts;

(d) Provisions should be made for clearing examinations in parts, in conformity with the modular pattern of courses;

(e) The practice of scaling marks of different subjects which are not at par may be adopted in determining the grades;

(f) Intensive training programmes will be organised for paper setters;

(g) Question banks will be developed to assist paper setters;

(h) A detailed marking scheme will be developed to ensure objectivity in scoring answer scripts;

(i) Innovative ideas like *open book examination*, diagnostic evaluation etc. may be experimented with;

(j) Separate certificates will be awarded showing the results of institutional evaluation and external examinations;

(k) The certificate of institutional evaluation may cover academic achievements as well as non-scholastic aspects.

(l) Attempts will be made to move towards a situation in which only those who teach will evaluate their students;

(m) Integration of evaluation with the process of teaching and learning will help diagnose the weaknesses and deficiencies in education. This diagnostic aspect will be utilised to develop remedial programme for weaker sections.

(n) Facilities will be provided in schools and colleges for maintenance of students' records to facilitate continuous institutional evaluation; and

(o) Programmes of training and orientation of teachers will give special attention to new evaluation methodologies, setting of question papers, measurement of performances, etc.

5. Emergent Emphases

(a) *Teacher assessments*

If one tries to analyse the philosophy underlying the New Educational Policy vis-a-vis Programme of Action there seems to be the growing faith in teacher on whom ultimately depends the quality of instruction vis-a-vis students' learning. More and more faith is being reposed on the class-room teacher whose assessment of his pupils are to be regarded more dependable than an external agent certifying their achievement. Teachers alone can illuminate the strength and weakness of their wards at regular intervals, being in constant touch with them. This provides the needed basis and facilitates remedial enrichment programme.

(b) *Diagnostic evaluation*

Another belief reflected in these recommendations is that evaluation should be used more as a teaching and diagnostic device than as a judgemental device. This it is believed would serve better the quality control function of evaluation. Thrust of reform has to be towards improvement of students' learning instead of merely measurement of their learning.

(c) *Mechanics of testing*

Need for training of paper setters and other functionaries reiterates the belief that technology of paper setting is well developed and demands trained paper setters, evaluators and

examiners. Like-wise decentralisation of the external examination for improving efficient administration, fair conduct of examination, and other measures recommended clearly indicate the need for improving mechanics of examinations.

(d) *Integrating evaluation*

The most significant indication is the stress on regarding evaluation as an integral part of the teaching learning process. Ecology of evaluation is highlighted, of course not to the desired extent, in the context of total learning milieu and social setting. How evaluation can be used as a feedback device to develop a dynamic equilibrium is reflected, though not in proper perspective.

(e) *De-emphasising external examinations*

We have been made to believe that the new education policy recognises teacher as real agent of examination reform and external public examinations must be replaced by institutional or teacher's evaluation. However, bold decision on doing away with external examination has yet not been taken. It is only de-emphasis of these examinations, which is recommended. Perhaps the administrators are afraid of maintaining standards at the hands of internal classroom teachers. Most probably they think of first establishing a well developed system of continuous comprehensive evaluation in schools before discontinuation of external examination.

Nevertheless, increasing demand and use of selection tests for admission to institutes of higher learning and for employment, is a sufficient evidence to do away with existing public examinations at the school stage to pave the way for teachers' evaluation. How soon the two incompatible philosophies, one of complete faith in teacher's evaluation and other of having continued faith in external examinations for certifying students achievement, are conceptualised in the form of teacher based assessment with quality control measures to maintain standards, depends on how soon we pin our complete faith in our teachers.

CHAPTER XIX

EMERGENT DEMANDS ON TODAY'S EVALUATOR

1. Introduction

In the previous chapters an attempt has been made to acquaint the readers with the concept and process of educational evaluation as well as the tools and techniques of evaluation. Discussion was limited to the evaluation of cognitive learning outcomes, interpretation of test scores, use of test-results and the changing emphases in the field of evaluation. The focus was the written examinations and their improvement at the school and public examination level. The clientele in view was the class teacher, the teacher educator and the inservice educator. Notwithstanding, the other aspects of evaluation and areas of evaluation which are emerging in this field, need special treatment. In fact during the seventies we find a clear drift away from measurement value to that of the pedagogical value of examinations. The reasons for this are quite a few. Some are educational, others are social while still others are administrative in nature. We shall try to identify special areas of thrust that need the attention of educators in general and of evaluators in particular. Three types of directions can be visualised for exploring emerging emphases. These may be based on the needs of learners, requirements of teachers and demands of the administrators.

After going through this chapter the reader is expected to

(a) identify the emerging demands related to the learner

e.g. self evaluation quizes, criterion referenced measures diagnosis and remediation, evaluation of the gifted etc.

(b) infer the teacher based demands in the context of affective and psychomotor outcomes evaluation as a teaching device, descriptive evaluation, question bank and institutional evaluation, etc.

(c) predict the administrator related demands like those of internal assessment, marking scaling and grading, curriculum and programme evaluation, open book examination etc.

(d) appreciate the special role of eventrator in tomorrow's school.

2. Emerging Demands

2.1. *Learner Based Demands*

(a) *Quizes for self-evaluation*

Various theories of learning stress the need for the development of positive self-concept so that the learner is well motivated for the tasks he is supposed to undertake. The evaluation system, therefore, must provide him rewarding experience in the form of success in a test. Feedback of failings should indeed be completely eschewed. What is required is the knowledge he should have about the adequacies and inadequacies in his learning so that he may make positive efforts to improve through self-evaluation. His performance need not be judged in terms of set norms or deviation from the normal. He should not be labelled as under achiever, a failure or a deviant. What is needed is his own introspection of his successes and failures. For this he may prefer self-instructional programmed material (1) like quiz cards which he can use as and when he likes to discover what he knows and what he does not. He might like to use review exercises given at the end of a chapter in a text-book. Class room questions by teachers and students' responses can also give some idea about ones' understanding of the subject. Other modes of self-evaluation can also be visualised to enable and guide the pupils to undertake self-analysis. It is therefore, suggested that use of Quizes may be taken advantage of as a self-evaluation technique.

(b) *Criterion-referenced measures*

The traditional evaluation system stresses the need for containing students' performance in terms of group performance i.e. class or school or a state norm. A learners' performance is judged against the performance of his peers or class mates. How good or poor his performance is, can be measured in terms of his deviation from the norm. This approach has lead to an erroneous belief that every learner can learn at the same rate and through same strategy. Like-wise it is assumed that every student has attained the intended objective. But this is not correct. The reality is that students with 50%, 40% and 33% marks have mastered the subject partially. Therefore, evaluation must be done in terms of criteria which may be stated in terms of intended instructional outcomes, or a performance standard regarded as expected learning. Moreover, the acceptable level in terms of mastery learning cannot be less than 70% to 80% mastery of the concepts. In other words, what is desired is mastery of the concepts beyond 70% by at least 70-75% of students. For this criterion referenced measures are needed as also the criterion referenced teaching model. (2) Thus emphasis on mastery learning and criterion referenced testing must be the focus of good teaching and testing. More discussion on this demands attention in a full-fledged chapter.

(c) *Diagnostic testing and remediation*

Diagnostic testing refers to the discovering of pupils' weaknesses for which there may be educational or non-educational causes. Among non-educational causes pupils' family background, physical defects, social adjustment problems etc. may be responsible for the poor performance. Among the educational causes there may be various reasons like poor educational background, non-mastery of basic concepts, lack of time provided for learning, defective instructional strategies etc. It is a well known fact that when weak students are promoted with 33% or 40% marks they are ultimately unable to get benefit from instruction and go on becoming weaker and weaker till they stagnate and drop out of the system. Unless proper diagnosis is undertaken and the reasons for their inadequacies in learning are identified, it becomes very difficult to provide

any remedial programme. Moreover, it is undemocratic to drag all the poor students along with the brighter ones in the instructional programmes and then declaring them as failures or under achievers without giving them an opportunity to improve. Care of the weaker section of students, especially at the elementary stage, is not only desirable but imperative in order to provide the needed remedial instruction and bring them up to the optimum level. (3)

This field of diagnosis and remediation has been almost neglected by most of our teachers. Therefore, in a heterogenous group as we have in most of the Indian schools, it is essential that special efforts be made to diagnose students' difficulties and then apply the needed correctives. Very few diagnostic tests have been developed and that too only in the field of Reading and Mathematics. Even at the doctorate level, we find very few studies in this field. A detailed discussion on the need and techniques of diagnostic testing and remediation is, therefore, a cogent demand of the learners to be met by teachers.

(d) *Evaluation for search of talent*

Inspite of the fact that a lot of work has been done to identify talented students, the fact remains that except for the national effort made at the N.C.E.R.T. level there is hardly any work being undertaken at the level of the states. West Bengal is of course already conducting a test for the award of scholarship for their students under the Jagdish Bose National Talent Search Scheme. Similar attempts need to be made in all the states in order to discover talented students who could be nurtured properly so that they may contribute to national development. A decentralised scheme of national talent search under the National Council of Educational Research and Training has been extended to the States from the year (1985-86) and for the first time a large scale effort at the State level was made by the respective States to identify their talented students through state level testing and for recommending candidates to the N.C.ER.T. stage testing. (4) Scheme of Navodaya Vidyalayas which envisages the opening of one school in each district of India by the end of seventh plan, is another landmark in the search for rural talent at the elementary stage and nurturing them in the

Navodaya Vidyalayas. This scheme is sponsored by Navodaya Vidyalaya Samiti under the Ministry of Human Resource Development, Government of India.

There are various schemes and scholarships which are awarded in various countries and the criteria of selection varies from the very subjective estimates like the teacher's assessment to that of very objective measures like those of a standardised tests. Nevertheless, at the school level what special measures are taken by the teachers in general to identify such students is yet to be seen. It would perhaps be appreciated if teachers are acquainted with the various types of assessment procedures for identifying such students and at the same time are provided guidance as to how enrichment programmes for such students may be given. It would, therefore, be in the national interest if the evaluators make a deliberate attempt to identify such students through well designed mental ability tests, aptitude tests and creativity tests to spot out talent. This area of assessment of creative abilities, therefore, needs to be discussed in more detail.

2.2. *Teacher-based Demands*

A lot of literature is available on evaluation of cognitive outcomes of learning. However, assessment of affective and psychomotor outcomes has not yet received the attention needed. It is yet to be realised by teachers that evaluation can be used for better learning and also for improvement of the teaching-learning process. Teachers have to appreciate that assessment of students' learning depends on types of inputs and the processes that are used and, therefore, any evaluation programme must be judged in terms of both the product and the process of learning. How examinations can be managed at the school level needs a well-thought out plan for evaluating students in schools? Likewise, the need for a good question bank cannot be over-emphasised. Keeping in view the teachers' limitation we may, therefore, like to focus the attention of the teachers to such areas.

(a) *Assessment of affective outcomes*

Affective outcomes of learning refer to the development of such non-cognitive behaviours which are mostly developed unintentionally, like appreciation, habits, interests, attitudes, values etc. Such outcomes of learning are developed incidentally, informally and non-formally and it is seldom that teachers make special effort for development of such learning outcomes. Because of their intangible nature, teachers have also not realised the necessity of assessing such outcomes. That is why even researches in the area of assessment on affective outcomes are not very many. It is only the highly technical experts who have devised instruments like attitude scales, interest inventories etc. that are usable only by experts. More functional and usable techniques in the class room need to be stressed. Techniques like observation, inquiry and analysis can be used. Taxonomies of educational objectives by Krathwhol (5) Hannah (6) in the area of affective domain are available but not much empirical evidence is available which can be used as a framework. Therefore, there is a need to explore the possibility of making assessment of affective outcomes of learning more objective with which teachers must be acquainted.

(b) *Assessment of Psychomotor Outcomes*

In contrast to the cognitive domain, work in the field of psychomotor domain has not been undertaken so seriously. There is hardly any study in India at the Ph.D. level. However, quite a few taxonomies of psychomotor domain have been advocated by experts like Simpson, Harrow, Alles, Dave and Hannah. Inspite of these taxonomies there appears to be not much empirical evidence available on the assessment of psychomotor outcomes. Even teacher educators and researchers in this field have yet to appreciate the urgent need for assessment of practical work. Areas like practical work in sciences, socially useful productive work, physical education, home science, etc. need the immediate attention of evaluators who could provide the needed framework for evaluation of students' learning in areas which involve psychomotor activities. The development of performance tests and ensuring the needed objectivity and reliability in marking is really a very challenging task for the

evaluators. Author himself has done some work in this field and it is reflected in a monograph (7) published by N.C.E.R.T. It hardly needs any testimony that in the area of practical work in all such subjects subjectivity of marking reigns supreme. There are a number of issues involved especially relating to the conduct of examination in these subjects and the quality of tools and techniques which are in use. It is, therefore, desirable that today's teacher must get an insight into the various problems and possibilities in the assessment of practical work in various subjects.

(c) *Evaluation for Better Teaching and Learning*

Almost every teacher knows that evaluation is a process of collecting evidences about pupils' learning and as such it concerns the various types of data gathering devices. The emphasis, therefore, is on the improvement of the reliability and validity of such evaluation instruments which are used for judging students' performance. The role of evaluation in making judgement is well known. These judgements are normally made in terms of group performance and are used for certifying students' achievement on the basis of which they are sometimes classified into various categories. Evaluation as a teaching device or as an experience for better learning is yet to be realised by the practising teachers. In fact, very few teachers are cognizant of the fact that evaluation can be used for improving students' learning besides measuring their achievement. How evaluation at the pre-instructional stage can help to ascertain the entry behaviour of students is not well known. The need for determining adequacies or inadequacies in the entry behaviour of the learner is essential for placement of students in the learning continuum. Therefore, evaluation at the pre-instructional stage helps the teachers to provide the basis for remediation before taking up developmental teaching.

Likewise, how many teachers are aware of the fact that evaluation in the form of class room questioning can go a long way in developing those very abilities which are tested through those very quetions? Well-structured and thought provoking questions used in the class help, in due course, to develop in students, the abilities of interpretation, analysis, hypothesis,

inference etc. (8) Therefore, the use of test material in teaching must be highlighted in any evaluation system. At the post instructional stage, when questions or a unit test is used, it provides data, which would help the teacher to know about adequacies and inadequacies in his teaching strategies. As emphasised in one of the previous chapters on objective based testing, the use of various levels of questions can be used to test various objectives which when used in class room teaching become the learning experience for the pupils who ultimately tend to put similar questions. Therefore, evaluation as a teaching device should also be realised by our teachers in order to improve their questioning ability. This aspect of evaluation also demands further elaboration.

(d) *Towards Descriptive Evaluation*

The most commonly used approach to testing is the psychometric approach which by and large aims at measurement of the product of learning. This approach is a statistical approach that attempts to measure students' learning in terms of gaps or congruence between the intended and the actual outcomes. Its focus is on precision and prediction. However, it totally ignores the social milieu in which learning takes place. The product of learning in fact, is the outcome of context, inputs and the process of teaching and learning. Every pupil learns within the given social setting and the facility he gets in terms of time, facilities available and the quality of instruction he receives. Thus, any product of learning or the impact depends on all these aspects. Unless evaluation takes into consideration all such variables and factors that throw light on pupils' achievement it will not be possible to pass any judgement about his product of learning. This sort of evaluation which emphasises more description and interpretation of data and explains the outputs of learning in terms of a given context, inputs and processes is termed as an illuminative evaluation. (9) This approach to evaluation needs the attention of the teachers and evaluators and requires separate treatment.

(e) *Management of Evaluation in Schools*

A teacher may be well acquainted with techniques of construc-

tion of various evaluation instruments, yet he may not be able to manage the evaluation programme properly due to the lack of insight in the relationship of various processes and the time targeting or scheduling of various activities. What type of tests should be used, at what time and for what purpose needs good planning. How are the test results to be recorded and interpreted for proper comparison of students' performance? How should judgements be formed and of what kind? What types of decisions may be taken regarding promotion, classification of students, diagnosis and remedial programmes? How should results be reported and communicated to the consumer? How much time should be devoted to testing in relation to instruction? How can students' performance in various schools or of various sections within the same school be compared? All this requires not only a broad conceptual framework of an evaluation system but also a systematic planning and scheduling of activities related to evaluation at various points of time for judging students' learning in various subjects. Management of school examinations is, therefore, an area that needs more work. (10)

(f) *Teachers' Question Bank*

Despite continuous organisation of various workshops for training of teachers in item writing, the fact remains that most teachers are not able to develop good questions, testing different abilities. Moreover, they do not have sufficient time and expertise to develop a good question bank in their subjects. It is true for most of teachers even in developed countries and that is why organisations like Educational Testing Services, and the National Foundation of Educational Research have taken up the responsibility of preparing quality materials in various subjects. If a good question bank is available with the teacher, he would be in a much better position to make use of that material and analyse his pupils' performance with a view to providing them better guidance for improvement of their learning. What type of question bank can a teacher prepare? What could be the strategies for preparing such a bank? How can coding of items be done? How can the question bank be used effectively? These are some of the issues which need to be discussed in more

detail. In fact, a framework for a teacher's question bank and its blue printing needs to be developed before taking up the process of item writing. A try-out of the items is essential to know the item parameters. It is, therefore, desirable that detailed scheme of question bank may be worked out at the institutional level and orientation for use of this material may be undertaken for the benefit of teachers. Need for educational testing service in this context cannot be over emphasised as reflected in N.C.E.R.T. (1985) National Curriculum Framework. (11)

2.3. *Administrators-based Demands*

It is quite obvious that a number of malpractices have crept in most of our examinations particularly in the public examinations conducted by various boards of school education. It is, therefore, necessary that some alternative measures may be thought of like those of the open-book examination. Similarly maintaining of standards has become almost difficult because of great diversity in the curriculum standard. What is needed is not only to ensure uniformity of standards all over India but also ways and means to make the scores comparable by different statistical measures like scaling. For quality control, it is necessary that more and more emphasis may be laid on evaluation. At the same time, an attempt may be made to evaluate the various curricula for renewal purpose. Similarly, programme evaluation must find a place in the total system of evaluation so that one should be able to see the congruence or the discrepancy between the intended and the observed results. A new programme of reform needs to be implemented in a phased manner by different agencies, like N.C.E.R.T, S.C.E.R.Ts, State Evaluation Units and State Institutes of Education that can play a significant role in the implementation of improvement programmes. Besides, research in examination must be an on-going process to see the effectiveness of the various evaluation devices as well as strategies of evaluation.

(a) *Internal Assessment and Continuous, Comprehensive Evaluation*
The need for internal assessment in the schools system has been

widely appreciated by all educationists. However, administrators have quite a few reservations regarding its effective implementation in schools as well as at the public examination level. There was a lot of supportive evidence of internal assessment programme during the 60s and 70s but for various reasons the role of internal assessment as a reliable method of judging pupils' performance in growth has been questioned. Even studies on internal assessment regarding attitudes of teachers are not favourable. At the public examination, internal assessment when shown separately in the certificate has created some problems for the teachers. When internal assessment scores are combined with the external assessment it leads to some academic problems of interpretation. Non-cognitive areas like attitudes, interests, habits, cocurricular activities etc. which have an important place in internal assessment do not find place in the public examinations, despite their place in the scheme of internal assessment. As such, the sanctity of internal assessment is sacrificed at the altar of practicability.

Should there be internal assessment at all stages of school education? Should internal assessment be based on only cognitive areas of development? Should marks on internal assessment be combined with those of external assessment? What should be the criteria of internal assessment? How can subjectivity in internal assessment be minimised? Do the teachers have a positive attitude towards internal assessment? These are some of the many more issues related to the scheme of internal assessment. Even the experience of NCERT working with States like Rajasthan (12) and Tamil Nadu (13) for a Comprehensive Scheme of Internal Assessment has not yielded the desired results. It has created more problems than it has solved. There are studies which clearly indicate the extreme subjectivity of teachers in assessing their students internally. There are instances of teachers even at the college level who would not like to undertake internal assessment for reasons which are more administrative than academic. Therefore it is necessary that some sort of survey of internal assessment practices be undertaken and problems identified, specially those relating to subjectivity in marking. How internal assessment can be made a practical preposition to provide more valid and reliable

evidence about pupils' growth, is a challenging task for the evaluators. Therefore, there is need for an incisive discussion on this major issue related to continuous, comprehensive evaluation.

(b) *Open Book Examinations*

It is an open secret that because of the various ills of the existing agencies and the malpractices with which these examinations are associated alternative measures have to be found out. A lot of debate has already gone into the experiment of open book examinations (14). Does the open book examination envisage the availability to students of all sorts of reference material or only a text book? Can the open book examination be used for testing all types of instructional outcomes? Can the usual type of questions be used or does it demand different expertise? Would it increase subjectivity in marking, because of the openness of questions? What would be the position of objective type questions in this system? Will the results of students on such an examination be comparable to that of those who take the existing examination? These and many more such questions need to be identified and discussed. To examine the feasibility of this system different tools and techniques may be explored and a discussion of their validity, reliability, and practicability will have to be initiated.

(c) *Marking, Scaling and Grading*

These three concepts are very much related and need to be understood properly in order to compare, maintain and set standards. What are the various marking systems? What are the various methods of grading? What are the various techniques of scaling? Is grading more reliable than numerical marking? Can scaling be justified? These are some of the many problems associated with the grading and scaling of results. These problems involve a lot of statistical handling and are not within the reach of the ordinary teacher. Nevertheless, what is needed is whether they are able to make the results or the scores more meaningful. Other statistics can be worked out even with the help of calculations under the guidance of experts. Since there is a lot of variation in the standards of question paper in

various subjects it is necessary that calibration of marks be undertaken to make the results more comparable. This area also needs more technical discussion (15).

(d) *Curriculum and Programme Evaluation*

It hardly needs any testimony that the quality of curriculum determines the standard of education. What are the curriculum components and how can each of them be evaluated? Objectives, content, methodology and outcomes of learning are basically the four major components which need to be evaluated for the renewal of curricula. Each of these components can be evaluated at various stages of curriculum development like the planning, development, implementation and quality control stages. For each of the components we will be able to use reflective evaluation, formative evaluation and summative evaluation at various stages of the curriculum development process. Likewise, when any programme is to be evaluated we have to evaluate its context, inputs, processes, products and the impact. Accordingly we have to resort to context evaluation, input evaluation, process evaluation, product evaluation and impact evaluation. (16) In fact, pupil evaluation, curriculum evaluation and programme evaluation are to be integrated as pupils' evaluation is sub-set of curriculum evaluation which in turn is a sub-set of programme evaluation. An evaluator must get a thorough acquaintance with this sort of integrated evaluation approach.

(e) *Research in Evaluation*

Research deals with finding out solutions for problems by providing evidences on the effectiveness of the various strategies involved. It not only provides evidences but it also highlights a number of inadequacies and adequacies in students' learning as well as in the teaching-learning strategies. It also provides data on the basis of which new hypotheses are formulated. Research on examination may relate to the validity and reliability of the instruments. It may relate to the mechanics of examination, economics of examinations or ecology of examinations. There is dire need for getting an idea into the various types of researches which have been undertaken in this field and to identify new

areas on which further research is needed. Likewise researches in curriculum evaluation, programme evaluation and institutional evaluation could also be undertaken to get empirical evidences about the efficacy of the curricula and programmes. It is, therefore, essential that our clientele get an insight into the various areas of researches that have been undertaken in the past and those which need to be undertaken to support or validate the various reform programmes.

3. Evaluators of Tomorrow

From the foregoing discussion, it is quite obvious that proper evaluation is not everybody's cup of tea. It needs a lot of insight into process of evaluation. It needs very systematic planning to identify the various issues relating to evaluation so that like a plaintiff, an evaluator can make out a good case. He is also required to defend his case when specially adverse judgements are made. He should be able to justify his stand in accordance with his terms of reference. Not only has he to advocate his view-point regarding the types of judgements he made but also why he made them. He might have to plead for self-evaluation in one case, norm-referenced evaluation in another case or criterion-referenced evaluation in still another case when a programme or a curriculum is to be evaluated. When there are conflicting views about the evaluation reports about implementation it is difficult for the administrators, he has to play the role of an arbitrator between the two parties and enable them to come to a consensus.

Lastly, he is supposed to pass clear cut judgements on the basis of evidences collected from various sources. He has to form his own judgements and take decisions accordingly. Whether a student should be promoted or not? Whether a student should be allowed to proceed to the next unit of learning? Whether a student should be given first class or second class? Whether a given curriculum is better than the other competing curriculum? Whether the results of my class are better than that of the other class? These are the types of situations where an evaluator has to exercise his own judgement and take decision as a judge.

Therefore, the job of an evaluator as a plaintiff, as a respondent, as a pleader, as an arbitrator and as a judge, must be appreciated by the evaluators of today. We would like to discuss these different roles in more details in a separate volume.

Ultimate objective of all evaluation is to improve student's learning through self-evaluation.

REFERENCES

CHAPTER I

1. DuBois, P.H. (1970): *A History of Psychological Testing*, Boston, Allyn and Bacon, Inc.
2. Garrett, H.E. (1951): *Great Experiments in Psychology*, Third Edition, Appleton, Inc. New York, pp. 175-81.
3. Russel, Charles (1930): *Standard Tests*, Ginn and Co., Boston, pp. 14-15.
4. Knight, Edgar, W. (1940): *Twenty Century Education*, Ginn and Co., Boston, pp. 52-53, 62.
5. Lang, Albert, R. (1930): *Modern Methods in Written Examinations*, Houghton Mifflin Co., Boston, pp. 2-3.
6. Cladwell Otis, W. and Courtis, Stuart, A. (1923): *Then and Now in Education 1845-1923*, World Book Co., Yonkers, New York, Chapters 1, 3.
7. Mann, Horace, (1845): *Common School Journal*, Vol. VII, No. 19.
8. Chadwick, E.B. (1864): Statistics of Educational Results, *The Museum*, A Quarterly Magazine of Education, Literature and Science, 3: 480-84; January.
9. Ayres, Leonard, P. (1918): The Measurement of Educational Results, *17th Year Book of National Society for Study of Education*, Part-II Public School Publishing Co., Bloomington, III, p. 11.
10. Scates, D.E. (1947): Fifty Years of Objective Measurement and Research in Education, *Journal of Educational Research*, 41: 241-64.
11. Thorndike, E.L. (1904): *An Introduction to the Theory of Mental and Social Measurement*, Teachers' College, Columbia University, New York.
12. Stone, Cliff, W. (1908): *Arithmetic Abilities and Some Factors Determining Them*, No. 12, Teachers College, Columbia University, New York.
13. Thorndike E.L. (1910): "Handwriting", Teacher's College Record 11: 83-175, March.
14. McCall, W.A. (1920): A New Kind of School Examination, *Journal of Educational Research*, 1: 33-46.
15. Ruch, G.M. (1924): *The Improvement of the Written Examinations*, Scott, Foresman and Co., Chicago.
16. Tyler Ralph, W. (1938): The Specific Techniques of Investigation: Examining and Testing Acquired Knowledge, Skill and Ability, *37th Year Book of National Society for the Study of Education*, Part II, Public

School Publishing Co., Chapter 29.
17. Kearney, Nolan, C. (1953): Elementary School Objectives, Report of the Mid-Century Committee on Outcomes in Elementary Education, Russel Sage Foundation, New York.
18. Smith, Eugene, R., Tyler, Ralph, W., et al. (1942): *Appraising and Recording Student Progress* Harper and Brothers, New York.
19. Bloom, B.S. et al. (1956): *Taxonomy of Educational Objectives*, Handbook-I Cognitive Domain, David McKay, New York.
20. Krathwhol, D.R. et al. (1964): *Taxonomy of Educational Objectives*, Handbook-II, Affective Domain, Longman, New York.
21. Simpson, E.J. (1964): *The Classification of Educational Objectives*, Handbook-III, Psychomotor Domain, University of Illinois.
22. Harrow, A.J. (1972): *A Taxonomy of the Psychomotor Domain*, David McKay, New York.
23. Despatch, Sir Charles (1854): Wood's Despatch.
24. Hunter, Sir, W.W. (1882): Report of Indian Education Commission (1882-83) Calcutta, Government Printing.
25. Indian Universities Commission (1902): Report of Indian Universities Commission.
26. Curzon, Lord (1904): Resolution on Indian Educational Policy Government of India.
27. Government of India Resolution (1913): Resolution on Indian Educational Policy.
28. Saddler, Michael (1919): Report of the Calcutta University Commission (1917-19).
29. Hartog, P. (1929): Report of the Indian Statutory Commission (Hartog Report).
30. Sapru Committee (1934): Report of the Sapru Committee, Government of U.P.
31. Sargeant Report (1944): *Post-War Educational Developments*, Indian Government Printing.
32. Radha Krishnan, S. (1950): Report of the University Education Commission, Government of India, New Delhi.
33. Mudaliar, A.L. (1952-53): Report of the Secondary Education Commission Government of India.
34. Narendra Deva Acharya (1953): Report of U.P. Committee on Reorganisation of Secondary Education.
35. A.I.C.S.E. (1955): Report of the First Meeting of All India Council for Secondary Education, Government of India.
36. D.E.P.S.E. (1957): Report of the First Conference of the Chairman and Secretaries of Boards of Secondary Education D.E.P.S.E., New Delhi.
37. D.E.P.S.E. (1958-61): Reports of 2nd, 3rd and 4th Conferences of Chairmen and Secretaries of Boards of Secondary Education D.E P.S.E., New Delhi.
38. Nag Chaudhary (1969): Report of the Committee on Reorganisation of N.C.E.R.T.

39. D.E.P.S.E. (1963-67): Reports of 5th, 6th, 7th and 8th Conferences of Chairmen and Secretaries of Boards of Secondary Education. D.E.P.S.E., New Delhi.
40. Mudaliar, A.L. (1952-53): Report of the Secondary Education Commission, Government of India.
41. Ministry of Education (1968): National Policy of Education, Ministry of Education, Government of India.
42. Rao, V.K.R.V. (1971): Report of the Committee on Examinations C.A.B.E., New Delhi.
43. University Grant Commission (1976): *Examination Reform—A Plan of Action*, U.G.C., Zafar Marg, New Delhi.
44. N.C.E.R.T. (1975): *The Curriculum for the Ten-Year School—A Framework*, N.C.E.R.T. New Delhi.
45. N.C.E.R.T. (1976): Higher Education and its Vocationalisation—Approach Paper, N.C.E.R.T., New Delhi.
46. Ministry of Human Resource Development (1986): National Policy on Education—Ministry of H.R.D., Department of Education, New Delhi.

CHAPTER II

1. Lindquist, E.F. (ed.) (1951): *Educational Measurement*, American Council of Education, Washington, D.C., pp. 533-40.
2. Popham, James, W. (1981): *Modern Educational Measurement*, Englewood Cliffs, N.J. Prentice Hall.
3. TenBrink, Terry, D. (1974): *Evaluation—A Practical Guide for Teachers*, McGraw Hill Book Company, New York, p. 11.
4. Gronlund, N.E. (1971): *Measurement and Evaluation in Teaching*, 2nd ed., The Macmillan Company, New York.
5. Bloom, B.S., et al. (1956): *Taxonomy of Educational Objectives*, Handbook-I, Cognitive Domain, David McKay Co., Inc., New York.
6. Scriven, Michael (1967): "The Methodology of Evaluation", AERA Monograph series on Curriculum Evaluation, No. 1, Rand McNally & Co., Chicago.
7. Bloom, B.S., Hastings, J.T. and Madaus, G.F. (1971): *Handbook of Formative and Summative Evaluation of Student Learning*, McGraw Hill Book Co., New York.
8. Singh, Pritam (1985): Diagnostic Evaluation and Remediation for Effective Learning, N.C.E.R.T., New Delhi.
9. Mager, R.F. (1962): *Preparing Instructional Objectives*, Palo Alto Calif.
10. Hamilton, et al. (1974): *Classroom Research: A Cautionary Tale*, Research in Education, 11.
11. Trow, M. (1957): Comment on "Participant Observation and Interviewing: A Comparison", Human Organisation, 16(3) pp. 33-5 (In McCall and Simmons, 1969).

12. Hamilton, D. et al. (1977): *Beyond the Number Games*, Basing Stock and London, Macmillan Education.
13. Popham, W.J. (1981): *Educational Evaluation—Goil Attainment Models.*
14. Tyler, Ralph, W. (1949): *Basic Principles of Curriculum and Instruction.* The University of Chicago Press, First Britain Impression.
15. Scriven, Michael (1971): *Goal Free Evaluation*, Part-I, 11-71, N.I.E., 2-A.
16. Stenhouse, Lawrence (1975): *Humanities Curriculum Project—An Introduction to Curriculum Research and Development*, London, Heinmann.

CHAPTER III

1. Garrett, Henry, E. (1958): *Statistics in Education and Psychology*, Allied Pacific Pvt. Ltd., Bombay, Indian Ed., p. 27.
2. Garrett, Henry, E. (1958): p. 29.
3. Garrett, Henry, E. (1958): p. 31.
4. Garrett, Henry, E. (1958): p. 50.
5. Garrett, Henry, E. (1958): p. 69.
6. Scannel, D.P. and Tracy, D.B. (1975): *Testing and Measurement in the Classroom*, Houghton Mifflin Co., Boston, p. 190.
7. Scannel, D.P. and Tracy, D.B. (1975): *Testing and Measurement in the Classroom*, Houghton Mifflin Co., Boston, p. 192.
8. Garrett, Henry, E. (1958): p. 143.
9. Garrett, Henry, E. (1958): p. 372.
10. TenBrink, Terry, D. (1974): *Evaluation—A Practical Guide for Teachers*, McGraw Hill Co., New York, pp. 451-55.
11. Diedrich, P.B. (1964). *Short-cut Statistics for Teacher Made Tests*, Princeton: Educational Testing Services, pp. 19, 34-36.
12. Best, John, W. and Kahn James (1986): *Research in Education*, Prentice Hall of India Pvt. Ltd., New Delhi, 3rd Edition (Indian Ed.), p. 337.
13. Best, John, W. and Kahn, James, V. (1986): *Research in Education*, Prentice Hall of India Pvt. Ltd., New Delhi, 3rd Edition (Indian Ed.), p. 338.
14. Wolfe, J.M. (1971): A Simple Method of Calculating the Coefficient of Correlation, *Journal of Educational Measurement*,8: pp. 221-22.
15. Scannel, D.P. and Tracy, D.B. (1975): *Testing and Measurement in the Classroom*, Houghton Mifflin Co., Boston, pp. 204-206.
16. Spearman, C. Brown, William Brown (1910): *British Journal of Psychology*, 2: pp. 271-95, 296-322.
17. Richardson, M.W. (1939): *The Calculation of Test Reliability Coefficient Based on Rational Equivalence*, J. Edu., Psychology 30: pp. 681-87.
18. Saupe, J.L. (1961): Some Useful Estimates of the Kuder Richardson Formula Number-20 Reliability Coefficient, *Educational and Psycho-*

logical Measurement, 2: pp. 63-72.
19. Garrett, H.E. (1958): *Statistics in Education and Psychology*, Allied Pacific Pvt. Ltd., p. 350.
20. Lord, F.M. (1959): Tests of the Same Length Do Have the Same Standard Error of Measurement, *Educational and Psychological Measurement*, 19: pp. 233-39.

CHAPTER IV

1. Gronlund, N.E. (1981): *Measurement and Evaluation in Teaching*, 5th Ed., Macmillan Collier, Macmillan, New York, pp. 25-26.
2. Furst, E.J. (1958): *Constructing Evaluation Instruments*, David McKay Co., Inc., New York.
3. Kothari, D.S. (1966): Report of the Education Commission (1964-66), Ministry of Education, Government of India, Chapter I.
4. Mager, R.F. (1962): *Preparing Instructional Objectives*, Belmont, Calif: Fearon Publishers.
5. National Education Association Commission on Reorganisation of Secondary Education, *Cardinal Principles of Secondary Education*, U.S. Office of Education, U.S.A.
6. Ebel, R.L. (1966): *Measuring Educational Achievement*, Prentice Hall of India Pvt. Ltd. New Delhi, 1969-70 (Indian Reprint).
7. Guilford, J.P. (1967): 'Structure of Intellect' in *The Nature of Human Intelligence*, N.T.C., McGraw Hill.
8. Bloom, B.S. et al. (1956): *Taxonomy of Educational Objectives*, Hand-Book-I, Cognitive Domain, David McKay, New York.
9. Gagne, Robert, M. (1965): *Conditions of Learning*, Holt, Rinehart & Winston, New York.
10. Madaus, G.E. (1973): In "Evaluation in Education International Progress, *International Review Series*, Vol. I, No. 2, 1977.
11. Hannah, Larry, S. and Michaelis, John, U. (1977): *A Comprehensive Framework for Instructional Objectives: A Guide to Systematic Planning and Evaluation*, Addison-Wesley Publishing Co., London.
12. Singh, Pritam (1977): An Investigation into the Empirical Validity of Bloom's Taxonomy of Educational Objectives, Unpublished Thesis, Delhi University.
13. Singh, Pritam (1986): In 'Evaluation at the Elementary Stage', *A Book of Readings*, N.C.E.R.T., New Delhi.
14. Kranthwohl, D.R. et al. (1964). *Taxomony of Educational Objectives*, Handbook-II, Affective Domain, David McKay, New Delhi.
15. Simpson, E.J. (1966): *Classification of Educational Objectives: Psychomotor Domain*, University of Illinois.
16. Alles, J. (1967): Analysis of Psychomotor Aspects of Behaviour, Theoretical Constructs in Curriculum Development and Evaluation, Sri Lanka.

17. Dave, R.H. (1971): Report of International Conference of Educational Testing at Berlin 1967, University of London Press.
18. Harrow, A.J. (1972): *A Taxonomy af Psychomotor Domain*, David McKay, New York.
19. Hannah, Larry, S. and Michaelis, John, U. (1977): *A Comprehensive Framework for Instructional Objectives: A Guide to Systematic Planning and Evaluation*, Addison-Wesley Publishing Co., London.
20. Maclay, R.W. (1969): *Organisation of Practical Work and Teaching in Integrated Science*, Post College Course for Teachers, Sidney Teachers College.
21. Moore (1967): *A Comprehensive Framework for Instructional Objectives: A Guide to Systematic Planning and Evaluation*, Addison-Wesley Publishing Co., London.
22. Eisner, Elliot, W. (1967): Educational Objectives: Help or Hindrance: School Review, 75, Autumn, pp. 250-60.
23. Atkin, J.M. (1968): "Behavioural Objectives in Curriculum Design: A Cautionary Note," *The Science Teacher*, 35.

CHAPTER V

1. Gronlund, N.E. (1986): *Measurement and Evaluation in Teaching*, 5th Ed., Collier and Macmillan, New York, pp. 10, 292-94.
2. Good, Carter, V. (Ed.) (1959): *Dictionary of Education*, McGraw Hill-Book Company, Inc., New York.
3. Good, Carter, V. (Ed.) (1959): *Dictionary of Education*, McGraw Hill-Book Company, Inc., New York.
4. Singh, Pritam (1983): Evaluating Students in Elementary Schools, Department of Measurement and Evaluation, N.C.E.R.T., New Delhi.
5. Good, Carter, V. (Ed.) (1959): *Dictionary of Education*, McGraw Hill-Book Company, Inc., New York, p. 34.
6. TenBrink, Terry, D. (1975): *Evaluation—A Practical Guide for Teachers*, McGraw Hill Book Company, Inc., New York, pp. 11, 12.
7. Lindquist (Ed.) (1951): *Educational Measurement*, Washington D.C., American Council on Education, pp. 533-40.
8. Gronlund, N.E. (1986): *Measurement and Evaluation in Teaching*, 5th Edition, Collier, Macmillan, New York, pp. 118-121.
9. TenBrink, Terry, D. (1974): *Evaluation—A Practical Guide for Teachers*, McGraw Hill Book Company, Inc., New York, Chapter 2.
10. McCormick, Robert & James, Mary (1983): *Curriculum Evaluation in Schools*, Billing and Sons Ltd., Worcester, U.K.

CHAPTER VI

1. N.C.E.R.T. (1979): *Minimum Learning Continuum*, Publication No. 3,

Primary Curriculum Development Cell, N.C.E.R.T., New Delhi.
2. Gronlund, N.E. (1970): *Stating Behavioural Objectives for Classroom Instruction*, Macmillan, New York.
3. Block, James, H. and Anderson, Lorin, W. (1975): *Mastery Learning in Classroom Instruction*, Macmillan Publishing Co., Inc., New York.
4. Singh, Pritam (1974): Objective Based Teaching and Testing, Sample Teaching-Learning Units in Biology, Department of Measurement and Evaluation, N.C.E.R.T.
5. TenBrink, Terry, D. (1974): *Evaluation—A Practical Guide for Teachers*, McGraw Hill Book Co., New York, Chapter I.
6. Bloom, B.S. (1971): *Mastery Learning and Its Implications for Curriculum Development* In Confronting Curriculum Reform by E.W. Eisner, Life Brow and Company, Inc., New York, pp. 17-49.
7. Singh, Pritam (1986): Diagnostic Evaluation and Remediation Department of Measurement & Evaluation, N.C.E.R.T.
8. Block, James, H. & Anderson, L.W. (1975): *Mastery Learning in Classroom Instruction*, Macmillan Publishing Co., Inc., New York, pp. 33-38.
9. Traxler, A.E. (1959): Ten Essential Steps in a Testing Programme, Education, 79: pp. 1-6.
10. Singh, Pritam (1983): Evaluating Students in Elementary Schools, Theory into Practice, Department of Measurement and Evaluation, N.C.E.R.T., New Delhi, p. 69.

CHAPTER VII

1. Horrance Mann (1845): *Common School Journal*, Vol. VII, No. 19.
2. D.E.P.S.E. (1964): Fourth Conference of Chairmen and Secretaries of Boards of Secondary Education, New Delhi.
3. Hudson, Barrie (1973): *Assessment Techniques—An Introduction*, Methuen Educational Ltd., London, pp. 167-72.
4. Board of Secondary Education, Rajasthan & N.C.E.R.T. (1965-66): Scheme of Comprehensive Internal Assessment. Rajasthan Board of Secondary Education, Ajmer.
5. Schwartz, A. and Tiedman, Stuart (1957): *Evaluating Students' Progress*, Longmans' Green and Company, New York.
6. Singh, Pritam (1971): Score card. In Preparation and Evaluation of Textbooks in Biology. Appendix C-2, p. 107, N.C.E.R.T.
7. Best, John, W. and Kahn, James, V. (1986): *Research in Education*, 5th Ed. Prentice-Hall of India Private Limited, New Delhi, pp. 166-80.
8. Moreno, J.L. et al. (Eds.) (1960): *The Sociometry Reader*, The Free Press, New York.
9. Buros, O.K. (Ed.) (1959): *The Fifth Mental Measurement Year Book*, The Cryphon Press, Rorschach Review by J.J. Eysenck.
10. TenBrink, Terry D. (1974): *Evaluation –A Practical Guide for Teachers*,

McGraw Hill Company, New York, pp. 147-48.

CHAPTER VIII-IX

1. Webster, Naoh (1981): *New Webster's Dictionary of the English Language*, Delux Edition, p. 783, Delair Publishing Company, Inc., USA.
2. Singh, Pritam (1972): Dynamics of a Question, *Journal of Indian Education*, March Issue, N.C.E.R.T.
3. Monoroe, Walter, S. and Carter, Ralph (1923): The Use of Different Types of Thought Questions in Secondary Schools, University of Illinois, *Research Bulletin*, No. 14.
4. Weidemann (1941): Review of Essay Test Studies, *Journal of Higher Education*, 12: pp. 41-44.
5. Stalnaker, John, M. (1951): The "Essay Type of Examination", In *Educational Measurement*, by Lindquist, E.F. American Council of Education. pp. 495-530.
6. Stalnaker, John, M. (1951): The "Eassy Type of Examination", In *Educational Measurement*, by Lindquist, E.F. American Council of Education, pp. 495-530.
7. Harper, A.E. (1975): Researches on Examination, N.C.E.R.T., New Delhi, pp. 8-52.
8. Singh, Pritam et al. (1977): Report of Workshop on Developing Question.
9. Singh, Pritam et al. (1977). Report of Banks' Examination Reform Unit, N.C.E.R.T., New Delhi.

CHAPTER X

1. Ebel, R.L. (1965): *Measuring Educational Achievement*, Prentice-Hall of India Pvt. Ltd., New Delhi, (Indian Print), p. 151.
2. Ibid., p. 160.
3. Ibid., p. 185.
4. Ibid., p. 184.
5. Ibid , p. 186.
6. Ibid., p. 179.
7. Ibid., p. 181.
8. Ibid., p. 174.

CHAPTER XI

1. Ebel, R.L. (1966): *Measuring Educational Achievement*, Prentice-Hall

of India Pvt. Ltd., New Delhi, pp. 144-45.
2. Ibid., pp. 124-27.
3. Ibid., pp. 130-31.
4. Swineford Frances (1941): Analysis of a Personality Trait, *Journal of Educational Psychology*, XXXII, pp. 438-44.
5. Chronbach, L.J. (1942): *Journal of Educational Psychology*, XXX, in Ebel, R.L. (1966) p. 135.
6. Remmers, H.H. et al. (1967): *A Practical Introduction to Measurement and Evaluation*, (Indian Reprint) U.B.S., Delhi, pp. 253-55.
7. Ibid., pp. 260-61.
8. Hudson, B. (1973): Mathuen Educational Ltd., London, p. 92.
9. Ibid., pp. 95-98.
10. Ibid., p. 103.
11. Head, J.J. (1967): Flexibility in Interpretation of an O. Level Mark Scheme, *Educational Review*, 19, pp. 18-128.

CHAPTER XII

1. Ebel, R.L. (1965): Prentice-Hall of India Pvt. Ltd. (Indian Reprint) New Delhi, Chapter 10.
2. Lindquist, E.F. (1955): *Educational Measurement*, American Council of Education, Washington D.C., p. 568.
3. Gheiselli, E.E. (1953): *Journal of Applied Psychology*, 37 pp. 18-20.
4. Guilford, J.P. (1956): *Fundamental Statistics in Psychology and Education*, McGraw Hill Book Co., Inc., London, p. 436.
5. Nunnally, J.C. (1907): *Psychometric Theory*, McGraw Hill Book Co., New York.
6. Cureton, E.E. (1956): In Lindquist, Educational Measurement, American Council of Education, Washington D.C., p. 622.
7. Richardson, M.W. (1939): *The Calculation of Test Reliability Coefficients Based on Method of Rational Equivalence*, J. Edn. Psycholo, 30 pp. 681-87.
8. Thorndike, R.L. and Hagen, E.P. (1969): *Measurement and Evaluation in Education and Psychology*, 3rd Edn., New York, p. 192.
9. Guilford, J.P. (1957): *Fundamental Statistics in Psychology and Education*, McGraw Hill Book Co., Inc., London, p. 457.
10. Lindquist, E.F (1955): *Educational Measurement*, American Council of Education, Washington D.C., p. 577.
11. TenBrink, Terry D. (1974): *Evaluation—A Practical Guide for Teachers*, McGraw Hill Book Co., New York, p. 461.
12. Lindquist, E.F. (1942): *First Course of Statistics* (Rev. ed., Boston) Houghton Mifflin Company, p. 213.
13. Cureton, E.F., Thorndike, R.L. and Hagen (1951): "Validity" In *Educational Measurement*, ed. E.F. Lindquist, Washington D.C., American Council of Education, p. 265.

14. English and English (Dictionary).
15. Thorndike, R.L. and Hagen (1955): *Measurement and Evaluation in Education and Psychology*, 3rd Ed. New York, pp. 109-110.

CHAPTER XIII

1. Singh, Pritam (1984): Objective Based Teaching and Testing-Sample Teaching-Learning Units in Biology, N.C.E.R.T., New Delhi, pp. 9-11.
2. Singh, Pritam (1988): Unit Tests in Biology for Classes IX & X (Under Print) Department of Measurement and Evaluation, N.C.E.R.T., New Delhi.
3. Anderson, L.W. (1975): *Mastery Learning and Classroom Instruction*, pp. 26-29, MacMillan Publishing Co., Inc., New York.
4. Menon, Kamala (1980): Sample Unit tests in Geography, Department of Measurement and Evaluation, N.C.E.R.T., New Delhi.
5. Singh, Pritam (1986): Diagnostic Evaluation and Remediation for Effective Learning, Department of Measurement and Evaluation, N.C.E.R.T., pp. 30-36.
6. Srivastava, H.S. and Singh, Pritam (1977): Use of Test Material in Teaching, N.C.E.R.T., New Delhi.
7. Singh, Pritam (1973): Biology Part-I Science for Middle Schools, Test items, Department of Science Education, N.C.E.R.T., New Delhi.

CHAPTER XIV-XV

1. Bloom, B.S. et al. (1956): *Taxonomy of Educational Objectives*, Handbook-I Cognitive Domain, David McKay, New York.
2. Singh, Pritam (1975): In Monograph on 'Practical Examination in Science Subjects' by Pritam Singh, N.C.E.R.T., New Delhi, pp. 20-23.
3. Rajasthan Board of Secondary Education & N.C.E.R.T. (1967): Sample question papers in Biology. Rajasthan Board of Secondary Education.
4. Singh, Pritam (1988): Sample question paper in Biology for classes IX-X (Under Print).
5. Hudson, B. (1973): *Assessment Techniques—An Introduction*, Methuen Educational Ltd., London, p. 196.
6. Ibid., p. 188.

CHAPTER XVI

1. Singh, Pritam (1966): Improving Teaching and Testing in Biology, State Evaluation Unit, Himachal Pradesh, Solan, Shimla Hills, pp. 68-77.
2. Saupe, J.L. (1961): Some Useful Estimates of the Kudar-Richardson Formula-20 Reliability Coefficient-Educational and Psychological Measurement, 2, pp. 63-72.

3. McMorris (1972): Quoted from Monograph on Test and Item Analysis for Universities by V. Natrajan, A.I.U., New Delhi (1977).
4. Diederich, Paul (1960): *Short cut Statistics for Teacher Made Tests*, Princeton.
5. TenBrink, Terry, D. (1974): Spearman Brown Formula for Correcting Split-half Correction Coefficient. Quoted from, *Evaluation—A Practical Guide for Teachers*, McGraw Hill Book Co., New York, p. 457.
6. TenBrink, Terry, D. (1974): Kudar-Richardson Formula-21 Quoted from, *Evaluation—A Practical Guide for Teachers*, McGraw Hill Book Co., New York, p. 457.
7. Chronbach, L. (1951): Quoted from *Monograph on Test and Item Analysis*, by V. Natrajan A.I.U. Delhi (1977), p. 183.
8. Nuttal (1969): Quoted from *Monograph on Test and Item Analysis* by V. Natrajan, A.I.U., New Delhi (1977), p. 183.

CHAPTER XVII

1. N.C.E.R.T. (1975): The Curriculum for the Ten-Year School—A Framework, N.C.E.R.T., New Delhi.
2. Department of Curriculum & Evaluation, N.C.E.R.T. (1968): Sample Unit Tests and Sample Question Papers in Biology, Board of Secondary Education, Assam, Guwahati.
3. Singh, Pritam (1975): Evaluating Students at the Elementary Stage, Department of Measurement and Evaluation, N.C.E.R.T., p. 69.
4. Singh, Pritam (1977): Board of Secondary Education, Assam, Guwahati.
5. Singh, Pritam (1983): Illuminative Evaluation in Science, School Science, June Issse, N.C.E.R.T.
6. Singh, Pritam (1983): Criterion-Referenced Testing—A Monograph, N.C.E.R.T.
7. Singh, Pritam (1985): Research on Examinations – A Retrospect and Prospect. In Contemporary Issues in Public Examinations, N.C.E.R.T., pp. 101-31.

CHAPTER XVIII

1. Ministry of H.R.D., Govt. of India (1986): National Policy on Education—1988, Government of India, Ministry of H.R.D., Department of Education, New Delhi.
2. Ministry of H.R.D., Govt. of India (1986): National Policy on Education—1986, Programme of Action, Ministry of H.R.D., Department of Education, New Delhi.

CHAPTER XIX

1. Singh, Pritam (1974): Use of Quizes in Radio Broadcast. Paper read in UNICEF Workshop organised by N.C.E.R.T.
2. Scannell, D., Tracy, D.B. (1975): *Testing and Measurement in the Classroom*, Houghton Mifflin Company, Boston, p. 39.
3. Singh, Pritam (1985): Diagnostic Evaluation and Remediation for Effective Learning, D.M.E.S. & D.P., N.C.E.R.T.
4. N.C.E.R.T. (1986): National Talent Search Scheme, N.C.E.R.T., New Delhi.

4a. Navodaya Vidyalaya Samiti, Ministry of H.R.D. (1987): Navodaya Vidyalaya Scheme, Ministry of Human Resource Development, Department of Education, New Delhi.

5. Krathwhol, D.R. et al. (1964): *Taxonomy of Educational Objectives*, Hand Book II, Affective Domain, David McKay, New York.
6. Hannah, Larry, S., Michaelis, John, (1977): A Comprehensive Framework of Instructional Objectives: *A Gutde to Systematic Planning and Evaluation*, Addison-Wesley Publishing Co., London.
7. Singh, Pritam (1983): A Monograph on Improving Practical Examinations in Science, N.C.E.R.T., New Delhi.
8. Srivastava, H.S., Singh, Pritam (1977): Use of Test Material in Teaching, N.C.E.R.T., New Delhi.
9. Singh, Pritam (1983): Illuminative Evaluation in Science, School Science, June Issue, N.C.E.R.T., New Delhi.
10. Singha, H.S. (1986): Management of Examinations—A Model. In Evaluation at the Elementary Stage—A Book of Readings, pp. 103-14, by Pritam Singh (Edit.) N.C.E.R.T., New Delhi.
11. N.C.E.R.T. (1985): National Curriculum for Primary and Secondary Education—A Framework—N.C.E.R.T., New Delhi.
12. Department of Curriculum & Evaluation, N.C.E.R.T. (1967): Scheme of Comprehensive Internal Assessment. Board of Secondary Education, Rajasthan, Ajmer.
13. S.C.E.R.T., Tamil Nadu (1976): State Council of Educational Research and Training, Tamil Nadu.
14. Arora, P.N. (1986): Open Book Examinations, D.M.E.S. & D.P. N.C.E.R.T., New Delhi.
15. Srivastava, A.B.L. (1985): Scaling of Examination Results. In Contemporary Issues in Public Examinations, pp. 342-347, N.C.E.R.T., New Delhi.
16. Singh, Pritam (1985): Indian Journal of Education, N.C.E.R.T., New Delhi.

INDEX

A.A.A.S Plan, 57
Abbot Wood Report 1936-37, 12
Ability tests, 425
Accountability, 31
Acharya Narendra Dev Committee 1948, 12
Achievement
  definition of, 147
Achievement test
  aim of, 315
  standardised, 5
Acquisition assignments, 203
Adaptability
  in curriculum, 446-7
Adaptive routinisation, 130
Affective domain, 128-9
Affective evaluation, 37
Aims, 107-8
All India Council for Secondary Education 1955 (AICSE), 14-15
Alles, 130
Alternative response matching, 249
Analysis
  approaches for, 432-3
  criterion-referenced, 177
  illuminative approach of, 432-3
  norm-referenced, 176-7
  psychometric approach of, 432
  purposes of, 430
  self-referenced, 176
  see also Item analysis
Anecdotal record, 196
Appraisal, 35
  of activity Programme, 6
  definition of, 147
Assessment, 35
  method
    measures proposed for, 467-9
  modes of, 145
Assignment, 171
Associational clue, 274
Assumptions, 153-4
Attitude scale, 202
Averages
    measures of, 70-74

Behaviour Communicativeness, 115
Bhopal Seminar 1956, 15, 17
Biology
  objectives of, 383-4, 397-400
  question papers in
    sample, 401-2, 405-13
Bloom's taxonomy, 7, 124-5, 253-5, 383
Blue-print, 357, 370-71
  aim of, 365
  preparation of
    assumptions for, 357-8
    methodology, 358-63
    for question paper, 389
  questions based on
    framing of, 389-90, 403-4
  of unit tests
    sample, 372-3
  using of, 363-5
Boston examination project, 3

C.P. and Berar Committee, 12-13
Central Advisory Board of Education 1944, 12

Central Services Examinations, 60
Check list, 195, 299-330
Chronbach, 295, 340
co-efficient alpha, 436
Closed objectives, 118
Cognitive evaluation, 37
Column matching
varieties of, 297-8
Common core test, 425
Compositeness
in curriculum, 446
Compound matching, 298
Concurrent validity, 341
Conferences of Chairman and Secretaries of Boards of Secondary Education, 16-18
Constant alternatives
construction of
guidelines for, 286-9
format of, 282-3
forms of, 283-6
see also True-false items
Construct validity, 341-2, 344
implications of, 349-50
Content
objective, 115
validity, 342
Correctives, 182-3
Correlation
computing methods of, 87-92
concept, 83-93
tetrachloric, 87, 88t
Councellor's record, 172
Covert behaviour, 118-9
Criterion-referenced
judgements, 178
measurement, 46-56, 460, 476
Critical-incidence-technique, 109
Curricular validity, 342, 344
Curriculum, 152-3, 162
approach paper on, 444
assumptions, 445-6
key concepts in, 446-8
definition of, 152
development, 486
evaluation, 61-2
Curzon's resolution 1904, 10

Data, 67, 449
techniques for, 187-204, 205t, 206t
organisation
methods, 68-70
Decision-making, 179-80
Derived scores
types of, 79-82
Derived validity, 343
Descriptive evaluation, 481
Deviation score, 74-5
Diagnosis, 182
Diagnostic
evaluation, 42-4, 149, 151t, 472
studies, 109-10
testing, 458, 476-7
Diagram
abilities tested, 308-11
classification of, 307
varieties of, 307-8
Direct validity, 343
Disciplines, 111
Discrimination index
calculating formulas for, 439-41
Dispersion
measures of, 74-9
Distractors, 250

Education Commission 1964-66, 20-21
Educational evaluation
cognitive development and, 457-8
ecology of, 161-2
evidence-gathering, 458-9
formative function of, 457
generalisations of, 450-52
objective-based, 456
recommendations for, 464-73
scope of, 457-61
trends in, 452
bases of, 452-4
principles for, 455
Educational objectives
classification of, 122-4

Empirical validity, 342
implications of, 347-9
Enabling objective, 117
Entrance examination
development of, 9
Error, 154
copping up of
principles for, 154-7
Essay-tests, 220, 256-7, 324
classification of, 221
extended response type, 221
facility value of, 437-8
grading, 226
methods of, 227-8
improvement of, 223-6
restricted response type, 220-21
scoring objectivity
improvement suggested for, 229-32
scoring reliability
improvement, 228-9
varieties of, 220-2
Estimate
standard error of, 102
Evaluation, 33-4
aspects of, 8, 149-50, 151t, 448-50
as continuous process, 158-9
cooperative process, 159
tools and techniques for, 187-204, 205t, 206t
diagnosis and correctives in, 182
as dyanamic process, 160
ecology of, 161-2
Emerging demands on,
learner based, 457-87
teacher based, 457-87
administrator based, 457-87
integrative model of, 163
interpretation of, 177-80
measurement and, 34
relations with, 148-9
models of
category wise, 53-6
nature and scope of, 146-9
objective-centred, 158
principles, 155-7
process, 164-5
characteristics of, 157-61
examination reform and, 464-7
purpose, 165-8
steps involved, 165-86
reporting, 181-2
summarisation, 180-81
teachers and, 480-81
Evidence collection, 170-71
sub-stages for, 171-4
Evidence processing,
methodology of basis, 459
Examinations,
developments of, 3-29
origin of, 1-3
Exo-internal evaluation, 38
Experiential domain, 133-4
Experimental evaluation, 51-2
External evaluation, 38-9

Face validity, 342
Facilitating objective, 117-8
Facility index
calculating formulas for, 437-9
Factorial validity, 342
Feedback, 183-5, 447-8, 460
Formal evaluation, 39-40
Format clue, 274
Formative evaluation, 42-3, 149-50, 151t, 160
Frequency distribution (FD), 69-70

Gagne-Merrill Taxonomy, 125
Goal-free evaluation, 54-5
Goals
objectives and, 107-8
statements, 107
Grading, 485
analytico-synthetic method for, 229-30
Grammatical clue, 274-5
Graphic stimuli, 307-11
Guess-who technique, 201

Harper's formulas, 439-40
Hartog Committee 1929, 11

Heterogeneous responses, 277
Holistic method, 227-8
Homogeneous responses, 277
Hudson, 316
Hunter Commission 1882, 9

Illuminative evaluation, 52-3, 481
Indian University Commission 1902, 10
Informal evaluation, 40-1
Information gathering
  analysis of, 174
    modes, 175-7
  forms of, 174-5
  recording of, 175
  sources of, 172-4
  see also Data gathering
Inquiry
  techniques of, 198-203
Instructional objectives, 166
  formulation of, 167
Integrating evaluation, 473
  reforms proposed for, 469-71, 473
Integration
  in curriculum, 447
Intellect
  faces of, 123-4
Intellectual processes, 127
Internal assessment, 23, 483-4
Internal evaluation, 37-8
Interpretation, 177-8
Interval scale, 66
Interview, 200
Intrinsic validity, 342
Inventory, 199-200
Item analysis, 437-42
  use of, 442-3

Job analysis, 109
Johnson Upper Lower Index, 439-40
Judgements, 149
  formation of, 177-8
    bases, 178-9
  importance of, 180
  purpose of, 178-9

Key, 250
Knowledge objective, 384
Kothari Commission, 112
Kuder-Richardson formulas, 97-8, 433, 435-6
  simplified, 98

Language, 268-9
Learnability, 447
Logical evaluation, 50-51

Maclay's taxonomy, 131-2
Macro-evaluation, 59-60
Marking, 485-6
  conformity check in, 423-4
  expert judgement in, 420-22
  grading in, 423-5
  leniency vs. severity, 422-3
  moderation in
    measures for, 420-4
    methods, 424-7
    by standardisation, 425-6
    by statistical methods, 424-5
    steps for, 421-2
  scheme, 319-20
    preparation of, 390-91
    scoring key and, 378-9
Mastery test, 134-5
  combined key, 301
  single key, 300-301
Matching type items
  common errors in, 302
  constructing of
    steps in, 302
    suggestions for, 302-3
  design, 296-7
  usability of, 301-302
  varieties of, 297-301
McMorris formula, 434
Measurement, 32-3
  assumptions of, 153-4
  definition of, 147-8
  errors of, 95-6, 151
    causes, 96
  evaluation and, 34, 456
    relations with, 148-9
  levels of, 64-7

of relationships, 83-93
standard error of, 99-102, 330, 339
estimating formula, 100
Lord's method of calculating, 100-101
statistical concepts in
data, 67-8
data organisation, 68-70
derived scores, 79-82
statistical measures, 70-79
Measuring instruments,
reliability, 326-40
usability, 324-6
validity, 340-51
see also Tools
Mean, 70-71, 75
Median, 72-3
Mega evaluation, 59
Micro-evaluation, 58-60
Mode, 73-4
Moderating test, 425
Moderation, 417-8
in marking,
measures for, 420-4
methods, 424-7
meaning of, 417
post-examination, 420-27
purpose of, 417-8
of question papers
dos' and donts', 419-20
tasks implied in, 418-9
Morrison index, 438-9
Multiple choice type, 246
base of, 248
construction of
suggestions for, 267-81
criticism of, 247
format of, 248-51
matching items, 296-303
ordinary, 248, 250
potentialities of, 251-7
rearrangement type, 304-6
varieties of, 257-67
Mysore Conference 1965, 20

Nag Chaudhari Committee, 19
National Council of Education Research & Training (NCERT), 18-19, 383
approach papers of, 23-5
taxonomy of,
objectives, 128
Navodaya Vidyalayas, 477-8
Negative stem, 271
New curriculum *see* Curriculum, apppoach paper
New Delhi conference 1957, 16-17
New Education Policy
emphases of, 472-3
Nominal scale, 65-6
Non-formal evaluation, 40
Non-parametric data, 67
Norm referenced measurement, 44-5
Normal probability curve, 93-5
Nuttal formulas, 436, 439

Objective-based
evaluation, 53-4
testing, 383-4
Objective-type tests
facility value
calculation formula, 437
matching type items
multiple choice items
pictorial varieties
Rearrangement item
structured questions
see also constant alternatives
Objectives
classification of, 108
criterion-referenced, 134-6
desirability of
approaches, 109-10
dimensions of, 166-7
ability definition, 167
educational
classification of, 122-4
elements of, 115-7
educational, 142
formulation of, 167-8
goals and, 107-8
identification and stating of

levels for, 112-5
item bank and, 141-2
national level, 112
nature of, 105-9
selection criteria for, 111-2
sources and derivation of
empirical approach, 109-12
statement of,
common errors in, 119-21
taxonomies, 124-34
Observation
instruments of, 194-8
techniques of, 194-8
Open-book examination, 483-5
Open objectives, 118
Oral examinations, 35-6, 189-90
advantages of, 190-91
classification of, 190
history of, 1-2
purpose of, 190
validity and reliability of
improvement, 191-3
Ordinal scale, 66
Overt behaviour, 118-9

Paper setting
objectives in
predominance of, 391-6
see also Question papers
Parallel test, 425
Parametric data, 67
Pay-off evaluation, 55-6
Percentile
rank, 81-2
score, 82
Perception domain, 132-3
Performance
assessment methods of
measures proposed for, 467-9
Personality tests, 6
Physical tests, 2
Pictorial questions, 307-11
Placement devices, 202
Point score method, 227
Practical test, 36
Predictive validity, 343-4
limitations of, 347-9
Process evaluation, 57-8
Product evaluation, 56-7
Product-cum-process evaluation,
57-8
Programme of Action (POA), 467-72
Programme evaluation, 61, 486
Projective devices, 202-3
Projects, 170-71
Psychometric approach, 481
Psychomotor
assessment, 37
domain, 479-80
Push down principle, 125

Qualitative evaluation, 48-9
Quantitative evaluation, 47-8
Quartile deviation, 76-7
Question bank, 482-3
Question papers
blue print of, 389
designing of, 384-92
editing and consolidation, 390
framing of questions 389-90,
403-4
moderation of, 419-20
Question-wise analysis, 391, 401-2,
414-6
Questionnaire, 198-200
Questions,
characteristics of, 213-4
classification of, 218-20
dynamics of, 214-5
essay-type, 220-28
facets of, 209-11
forms of, 387-8
genesis of, 211-3
long-answer, 217-32
objective based, 217-8
paradigm of, 211, 212f
short-answer, 232-42
Quizes, 475

Radhakrishan Commission 1948-49,
13
Rankings, 69, 197
Rao Committee on Examinations
(1971), 21-3

Rating
method, 227-8
scales, 195-6
Ratio scales, 66-7
Rational evaluation, 50-51
Readiness testing, 168-9
Rearrangement items 304-6
Relationships,
measures of, 83-93
Reliability, 95-9, 326-40
coefficient, 330-2
computing methods, 97-9
definitions of, 327-9
estimation of
methods for, 336-8
evaluation of,
data needed for, 339
precautions to control, 228-9
of tests
formulas for, 433-6
validity and, 326, 329
Remediation, 168-9, 476-7
Reporting, 181-2
Research
evaluation, 486-7
Resolution of 1913, 10-11
Responses, 273-81
clues and, 274-6
length clue, 275-6
Responsive evaluation, 49-50
Review assignments, 204
Rice, J.M., 3-4

Saddler Commission 1917-19, 11
Sampling validity, 342
Sapru Committee 1934, 12
Sargent Report 1944, 12
Saupe formula, 98-9
Scales
types of, 65-7
Scaling, 485
Score cards, 197-8
Scores
confidence trait and, 293, 295
Scoring key
marking scheme and, 378-9
preparation of, 390-91
Scores
interpretability of, 325
true and error, 331-2
Secondary Education Commission, 1952-53, 13-4
Self evaluation quizes, 475
Self-referenced
judgements, 178
measurement, 46-7
Sentence completion clue, 276
Short-answer questions, 232-37
concept of, 232 3
see also Very-short-answer questions
Simpson's taxonomy, 129
Social distance scale, 201-2
Sociometry, 200-203
Spearman's-Brown
correction formula, 435
prophesy formula, 333-4, 337
Specification
meaning of, 121
Split-half method, 97, 99, 433, 435-6
Stem, 248
criteria for, 270-73
Standard deviation, 77-9
Standard error of measurement (SEM), 99-102, 330, 339
Standard scores, 79-80
Statistical analysis
descriptive, 67
inferential, 68
measures, 70-79
Statistics, 64
Statistical validity, 342-3
Structured questions, 311-20
characteristics of, 312-5
Subject matter
units of, 353
Summarisation, 180-81
Summative evaluation, 42-3, 150, 151t, 169

Talent search
examination, 60

scheme, 477-8
Taxonomies, 124-34, 383, 479
cognative domain, 124-8
psychomotor domain, 129-32
Teacher assessments, 472
Teacher's records, 172, 174
Teacher's training, 17
Teaching and testing
objectives' place in, 136-7, 138-9t, 140-41
Techniques, 187-8
inquiry, 193-203
observational, 194-8
testing, 188-93
TenBrink, 33, 94, 171-2, 203
Terminal objectives, 117
Test
definition of, 147
Test results, 429-31
Testing
devising situation for, 170-71
techniques of, 188-93, 204-205, 206t
Test-retest reliability, 97
Tests, 34-5, 429-30
formulas, 433-6
norm-referenced, 45
reliability, 326-40
requirements for good, 324-51
usability of, 324-6
validity, 340-51
see also, Examinations; unit tests; written tests
Tools, 188, 206
for analysis, 203-4
for enquiry, 198-203
for observation, 194-8
sociometric, 201-3
Transfer exercises, 204
True-false items, 283-296
see also constant alternatives

Unit concept, 353-4
Unit tests, 36?-6
blue-print of, 357-65
concept of, 354
duration of, 355
types of, 354-7
useful purposes of, 366-9
UP Committee on Reorganisation of Secondary Education 1953, 14
University Education Commission 1948-49, 13
Usability, 324-6

Validity, 340-51
definitions of, 340-41
implications of, 345-50
limitations
in use, 350-51
reliability and, 326, 329
types of, 341-3
categories of, 343-5
Values, 105-6
Variance, 75-6, 96-7
sources of, 327-8
Verbal clue, 275
Very-short answer questions
construction of
suggestions for, 239-42
types of, 238-9
Vocationalisation
approach paper on, 24-5

Wolfes adaptation methods, 90-92
Wood's despatch of 1854, 9
Written tests, 35-6, 145, 188-9, 382
advantages of, 3
origin of, 2-3
see also Examinations

ERRATA

Page	*Line*	
(vi)	10 from bottom	Read *school* in place of *cool*
(vii)	19 from top	Read *emphases* in place of *emphasis*
(xii)	03 from top	Delete the word 'which'
(xiii)	09 from bottom	Delete the word 'besides'
(xv)	04 from top	Read achieve*ment* in place of achieve*ments*
(31)	13 from top	Read *a* far cry in place of *of* far cry
(90)	13 from top	Read the formula without minus sign before 6
(95)	03 from top	Read the formula as $y = \frac{N}{\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$
(190)	03 from top	Read *Content* Subjects in place of *Continent* Subjects
(210)	Last line below the graph	Read *hours* in place of *tours*
(385)	13 from bottom	(ii) Subheading: Read Total marks in place of Total time
(435)	Last line	(v) K.R. Formula-20: Read in the formula Σpq in place of pq
(437)	03 from top	Read F.V. of an item $= \frac{\text{No. of students attempting the items correctly}}{\text{Total number attempting the items}} \times 100$
(465)	19 from top	Read article (8.24) in place of (8.21)
(467)	11 from bottom	Read Programme of *Action* in place of *actions*
(468)	17 from bottom	Read *Deemed* University in place of *Demand* University
(475)	11 from bottom	Read *evaluator* in place of *eventrator*

Please read *Evaluation in National Policy on Education (1986)* running headings on top of pages 465, 467, 469, 471 and